

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
9 January 2003 (09.01.2003)

PCT

(10) International Publication Number
WO 03/001976 A2

(51) International Patent Classification⁷: **A61B**

(21) International Application Number: PCT/US02/20336

(22) International Filing Date: 26 June 2002 (26.06.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/891,793 26 June 2001 (26.06.2001) US

(71) Applicant (for all designated States except US): **ISIS PHARMACEUTICALS, INC.** [US/US]; 2292 Faraday Avenue, Carlsbad, CA 92009 (US).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **ECKER, David, J.** [US/US]; 1041 Saxony Road, Encinitas, CA 92024 (US).

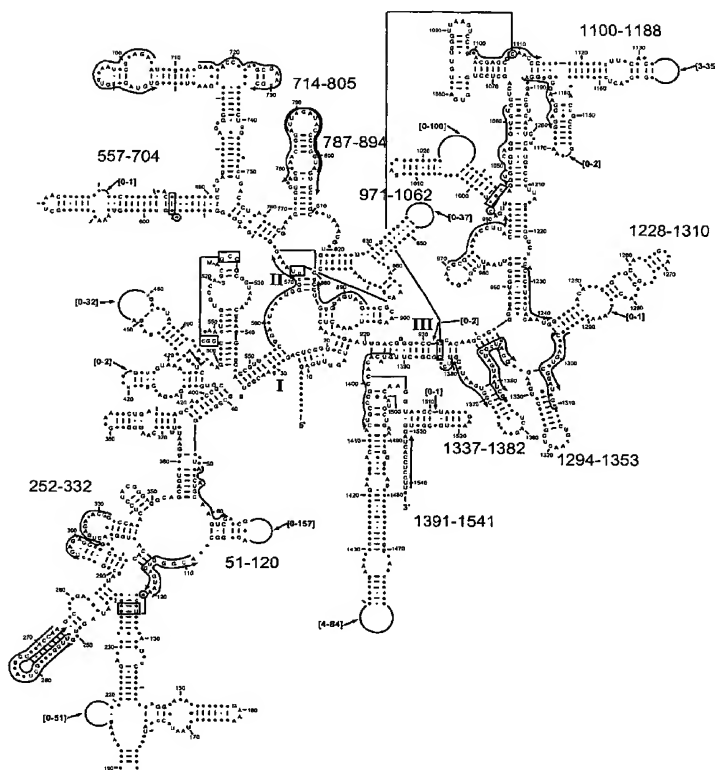
GRIFFEY, Richard, H. [US/US]; 360 Barsby Street, Vista, CA 92084 (US). **SAMPATH, Rangarajan** [IN/US]; 12223 Mannix Road, San Diego, CA 92129 (US). **HOF-STADLER, Steven, A.** [US/US]; 5014 Viewridge Way, Oceanside, CA 92056 (US). **MCNEIL, John** [US/US]; 437 Retaheim Way, La Jolla, CA 92037 (US). **CROOKE, Stanley, T.** [US/US]; 3211 Piragua Street, Carlsbad, CA 92009 (US).

(74) Agent: **LEGAARD, Paul, K.**; Woodcock Washburn LLP, 46th Floor, One Liberty Place, Philadelphia, PA 19103 (US).

(81) Designated States (*national*): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZM, ZW.

[Continued on next page]

(54) Title: A SECONDARY STRUCTURE DEFINING DATABASE AND METHODS FOR DETERMINING IDENTITY AND GEOGRAPHIC ORIGIN OF AN UNKNOWN BIOAGENT THEREBY



(57) Abstract: The present invention relates generally to the field of investigational bioinformatics and more particularly to secondary structure defining databases. The present invention further relates to methods for interrogating a database as a source of molecular masses of known bioagents for comparing against the molecular mass of an unknown or selected bioagent to determine either the identity of the selected bioagent and/or to determine the origin of the selected bioagent. The identification of the bioagent is important for determining a proper course of treatment and/or eradication of the bioagent in such cases as biological warfare. Furthermore, the determination of the geographic origin of a selected bioagent will facilitate the identification of potential criminal identity.

WO 03/001976 A2



(84) Designated States (regional): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— without international search report and to be republished upon receipt of that report

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

**A SECONDARY STRUCTURE DEFINING DATABASE AND METHODS
FOR DETERMINING IDENTITY AND GEOGRAPHIC ORIGIN OF AN
UNKNOWN BIOAGENT THEREBY**

5 STATEMENT OF GOVERNMENT SUPPORT

This invention was made with United States Government support under DARPA/SPO contract BAA00-09. The United States Government may have certain rights in the invention.

10 FIELD OF THE INVENTION

The present invention relates generally to the field of investigational bioinformatics and more particularly to secondary structure defining databases. The present invention further relates to methods for interrogating a database as a source of molecular masses of known bioagents for comparing against the molecular mass of an unknown or selected bioagent to determine either the identity of the selected bioagent, and/or to determine the origin of the selected bioagent. The identification of the bioagent is important for determining a proper course of treatment and/or eradication of the bioagent in such cases as biological warfare. Furthermore, the determination of the geographic origin of a selected bioagent will facilitate the identification of potential criminal identity. The present invention also relates to methods for rapid detection and identification of bioagents from environmental, clinical or other samples. The methods provide for detection and characterization of a unique base composition signature (BCS) from any bioagent, including bacteria and viruses. The unique BCS is used to rapidly identify the bioagent.

25

BACKGROUND OF THE INVENTION

In the United States, hospitals report well over 5 million cases of recognized infectious disease-related illnesses annually. Significantly greater numbers remain undetected, both in the inpatient and community setting, resulting in substantial morbidity and mortality. Critical intervention for infectious disease relies on rapid, sensitive and specific detection of the offending pathogen, and is central to the mission of microbiology laboratories at medical centers. Unfortunately, despite the recognition that outcomes from infectious illnesses are directly associated with time to pathogen recognition, as well as accurate identification of the class and species of

microbe, and ability to identify the presence of drug resistance isolates, conventional hospital laboratories often remain encumbered by traditional slow multi-step culture based assays. Other limitations of the conventional laboratory which have become increasingly apparent include: extremely prolonged wait-times for pathogens with
5 long generation time (up to several weeks); requirements for additional testing and wait times for speciation and identification of antimicrobial resistance; diminished test sensitivity for patients who have received antibiotics; and absolute inability to culture certain pathogens in disease states associated with microbial infection.

For more than a decade, molecular testing has been heralded as the diagnostic
10 tool for the new millennium, whose ultimate potential could include forced obsolescence of traditional hospital laboratories. However, despite the fact that significant advances in clinical application of PCR techniques have occurred, the practicing physician still relies principally on standard techniques. A brief discussion of several existing applications of PCR in the hospital-based setting follows.

15 Generally speaking molecular diagnostics have been championed for identifying organisms that cannot be grown *in vitro*, or in instances where existing culture techniques are insensitive and/or require prolonged incubation times. PCR-based diagnostics have been successfully developed for a wide variety of microbes. Application to the clinical arena has met with variable success, with only a few assays
20 achieving acceptance and utility.

One of the earliest, and perhaps most widely recognized applications of PCR for clinical practice is in detection of *Mycobacterium tuberculosis*. Clinical characteristics favoring development of a nonculture-based test for tuberculosis include week to month long delays associated with standard testing, occurrence of
25 drug-resistant isolates and public health imperatives associated with recognition, isolation and treatment. Although frequently used as a diagnostic adjunctive, practical and routine clinical application of PCR remains problematic due to significant inter-laboratory variation in sensitivity, and inadequate specificity for use in low prevalence populations, requiring further development at the technical level. Recent advances in
30 the laboratory suggest that identification of drug resistant isolates by amplification of mutations associated with specific antibiotic resistance (e.g., *rpoB* gene in rifampin resistant strains) may be forthcoming for clinical use, although widespread application will require extensive clinical validation.

One diagnostic assay, which has gained widespread acceptance, is for *C. trachomatis*. Conventional detection systems are limiting due to inadequate sensitivity and specificity (direct immunofluorescence or enzyme immunoassay) or the requirement for specialized culture facilities, due to the fastidious characteristics of this microbe. Laboratory development, followed by widespread clinical validation testing in a variety of acute and nonacute care settings have demonstrated excellent sensitivity (90-100%) and specificity (97%) of the PCR assay leading to its commercial development. Proven efficacy of the PCR assay from both genital and urine sampling, have resulted in its application to a variety of clinical setting, most recently including routine screening of patients considered at risk.

While the full potential for PCR diagnostics to provide rapid and critical information to physicians faced with difficult clinical-decisions has yet to be realized, one recently developed assay provides an example of the promise of this evolving technology. Distinguishing life-threatening causes of fever from more benign causes in children is a fundamental clinical dilemma faced by clinicians, particularly when infections of the central nervous system are being considered. Bacterial causes of meningitis can be highly aggressive, but generally cannot be differentiated on a clinical basis from aseptic meningitis, which is a relatively benign condition that can be managed on an outpatient basis. Existing blood culture methods often take several days to turn positive, and are often confounded by poor sensitivity or false-negative findings in patients receiving empiric antimicrobials. Testing and application of a PCR assay for enteroviral meningitis has been found to be highly sensitive. With reporting of results within 1 day, preliminary clinical trials have shown significant reductions in hospital costs, due to decreased duration of hospital stays and reduction in antibiotic therapy. Other viral PCR assays, now routinely available include those for herpes simplex virus, cytomegalovirus, hepatitis and HIV. Each has a demonstrated cost savings role in clinical practice, including detection of otherwise difficult to diagnose infections and newly realized capacity to monitor progression of disease and response to therapy, vital in the management of chronic infectious diseases.

The concept of a universal detection system has been forwarded for identification of bacterial pathogens, and speaks most directly to the possible clinical implications of a broad-based screening tool for clinical use. Exploiting the existence of highly conserved regions of DNA common to all bacterial species in a PCR assay would empower physicians to rapidly identify the presence of bacteremia, which

would profoundly impact patient care. Previous empiric decision making could be abandoned in favor of educated practice, allowing appropriate and expeditious decision-making regarding need for antibiotic therapy and hospitalization.

Experimental work using the conserved features of the 16S rRNA common to almost all bacterial species, is an area of active investigation. Hospital test sites have focused on “high yield” clinical settings where expeditious identification of the presence of systemic bacterial infection has immediate high morbidity and mortality consequences. Notable clinical infections have included evaluation of febrile infants at risk for sepsis, detection of bacteremia in febrile neutropenic cancer patients, and examination of critically ill patients in the intensive care unit. While several of these studies have reported promising results (with sensitivity and specificity well over 90%), significant technical difficulties (described below) remain, and have prevented general acceptance of this assay in clinics and hospitals (which remain dependent on standard blood culture methodologies). Even the revolutionary advances of real-time PCR technique, which offers a quantitative more reproducible and technically simpler system remains encumbered by inherent technical limitations of the PCR assay.

The principle shortcomings of applying PCR assays to the clinical setting include: inability to eliminate background DNA contamination; interference with the PCR amplification by substrates present in the reaction; and limited capacity to provide rapid reliable speciation, antibiotic resistance and subtype identification. Some laboratories have recently made progress in identifying and removing inhibitors; however background contamination remains problematic, and methods directed towards eliminating exogenous sources of DNA report significant diminution in assay sensitivity. Finally, while product identification and detailed characterization has been achieved using sequencing techniques, these approaches are laborious and time-intensive thus detracting from its clinical applicability.

Rapid and definitive microbial identification is desirable for a variety of industrial, medical, environmental, quality, and research reasons. Traditionally, the microbiology laboratory has functioned to identify the etiologic agents of infectious diseases through direct examination and culture of specimens. Since the mid-1980s, researchers have repeatedly demonstrated the practical utility of molecular biology techniques, many of which form the basis of clinical diagnostic assays. Some of these techniques include nucleic acid hybridization analysis, restriction enzyme analysis, genetic sequence analysis, and separation and purification of nucleic acids (See, e.g.,

J. Sambrook, E. F. Fritsch, and T. Maniatis, Molecular Cloning: A Laboratory Manual, 2nd Ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., 1989). These procedures, in general, are time-consuming and tedious. Another option is the polymerase chain reaction (PCR) or other amplification procedure which amplifies a specific target DNA sequence based on the flanking primers used. Finally, detection and data analysis convert the hybridization event into an analytical result.

Other not yet fully realized applications of PCR for clinical medicine is the identification of infectious causes of disease previously described as idiopathic (e.g. *Bartonella henselae* in bacillary angiomatosis, and *Tropheryma whippellii* as the uncultured bacillus associated with Whipple's disease). Further, recent epidemiological studies which suggest a strong association between Chlamydia pneumonia and coronary artery disease, serve as example of the possible widespread, yet undiscovered links between pathogen and host which may ultimately allow for new insights into pathogenesis and novel life sustaining or saving therapeutics.

For the practicing clinician, PCR technology offers a yet unrealized potential for diagnostic omnipotence in the arena of infectious disease. A universal reliable infectious disease detection system would certainly become a fundamental tool in the evolving diagnostic armamentarium of the 21st century clinician. For front line emergency physicians, or physicians working in disaster settings, a quick universal detection system, would allow for molecular triage and early aggressive targeted therapy. Preliminary clinical studies using species specific probes suggest that implementing rapid testing in acute care setting is feasible. Resources could thus be appropriately applied, and patients with suspected infections could rapidly be risk stratified to the different treatment settings, depending on the pathogen and virulence. Furthermore, links with data management systems, locally regionally and nationally, would allow for effective epidemiological surveillance, with obvious benefits for antibiotic selection and control of disease outbreaks.

For the hospitalists, the ability to speciate and subtype would allow for more precise decision-making regarding antimicrobial agents. Patients who are colonized with highly contagious pathogens could be appropriately isolated on entry into the medical setting without delay. Targeted therapy will diminish development of antibiotic resistance. Furthermore, identification of the genetic basis of antibiotic resistant strains would permit precise pharmacologic intervention. Both physician and

patient would benefit with less need for repetitive testing and elimination of wait times for test results.

It is certain that the individual patient will benefit directly from this approach. Patients with unrecognized or difficult to diagnose infections would be identified and treated promptly. There will be reduced need for prolonged inpatient stays, with resultant decreases in iatrogenic events.

Mass spectrometry provides detailed information about the molecules being analyzed, including high mass accuracy. It is also a process that can be easily automated. Low-resolution MS may be unreliable when used to detect some known agents, if their spectral lines are sufficiently weak or sufficiently close to those from other living organisms in the sample. DNA chips with specific probes can only determine the presence or absence of specifically anticipated organisms. Because there are hundreds of thousands of species of benign bacteria, some very similar in sequence to threat organisms, even arrays with 10,000 probes lack the breadth needed to detect a particular organism.

Antibodies face more severe diversity limitations than arrays. If antibodies are designed against highly conserved targets to increase diversity, the false alarm problem will dominate, again because threat organisms are very similar to benign ones. Antibodies are only capable of detecting known agents in relatively uncluttered environments.

Several groups have described detection of PCR products using high resolution electrospray ionization-Fourier transform-ion cyclotron resonance mass spectrometry (ESI-FT-ICR MS). Accurate measurement of exact mass combined with knowledge of the number of at least one nucleotide allowed calculation of the total base composition for PCR duplex products of approximately 100 base pairs. (Aaserud *et al.*, *J. Am. Soc. Mass Spec.*, **1996**, 7, 1266-1269; Muddiman *et al.*, *Anal. Chem.*, **1997**, 69, 1543-1549; Wunschel *et al.*, *Anal. Chem.*, **1998**, 70, 1203-1207; Muddiman *et al.*, *Rev. Anal. Chem.*, **1998**, 17, 1-68). Electrospray ionization-Fourier transform-ion cyclotron resistance (ESI-FT-ICR) MS may be used to determine the mass of double-stranded, 500 base-pair PCR products via the average molecular mass (Hurst *et al.*, *Rapid Commun. Mass Spec.* **1996**, 10, 377-382). The use of matrix-assisted laser desorption ionization-time of flight (MALDI-TOF) mass spectrometry for characterization of PCR products has been described. (Muddiman *et al.*, *Rapid*

Commun. Mass Spec., **1999**, *13*, 1201-1204). However, the degradation of DNAs over about 75 nucleotides observed with MALDI limited the utility of this method.

U.S. Pat. No. 5,849,492 describes a method for retrieval of phylogenetically informative DNA sequences which comprise searching for a highly divergent segment of genomic DNA surrounded by two highly conserved segments, designing the
5 universal primers for PCR amplification of the highly divergent region, amplifying the genomic DNA by PCR technique using universal primers, and then sequencing the gene to determine the identity of the organism.

U.S. Patent No. 5,965,363 discloses methods for screening nucleic acids for
10 polymorphisms by analyzing amplified target nucleic acids using mass spectrometric techniques and to procedures for improving mass resolution and mass accuracy of these methods.

WO 99/14375 describes methods, PCR primers and kits for use in analyzing preselected DNA tandem nucleotide repeat alleles by mass spectrometry.

WO 98/12355 discloses methods of determining the mass of a target nucleic
15 acid by mass spectrometric analysis, by cleaving the target nucleic acid to reduce its length, making the target single-stranded and using MS to determine the mass of the single-stranded shortened target. Also disclosed are methods of preparing a double-stranded target nucleic acid for MS analysis comprising amplification of the target
20 nucleic acid, binding one of the strands to a solid support, releasing the second strand and then releasing the first strand which is then analyzed by MS. Kits for target nucleic acid preparation are also provided.

PCT WO97/33000 discloses methods for detecting mutations in a target nucleic acid by nonrandomly fragmenting the target into a set of single-stranded
25 nonrandom length fragments and determining their masses by MS.

U.S. Patent No. 5,605,798 describes a fast and highly accurate mass spectrometer-based process for detecting the presence of a particular nucleic acid in a biological sample for diagnostic purposes.

WO 98/21066 describes processes for determining the sequence of a
30 particular target nucleic acid by mass spectrometry. Processes for detecting a target nucleic acid present in a biological sample by PCR amplification and mass spectrometry detection are disclosed, as are methods for detecting a target nucleic acid in a sample by amplifying the target with primers that contain restriction sites and tags, extending and cleaving the amplified nucleic acid, and detecting the presence of

extended product, wherein the presence of a DNA fragment of a mass different from wild-type is indicative of a mutation. Methods of sequencing a nucleic acid via mass spectrometry methods are also described.

WO 97/37041, WO 99/31278 and US Patent No. 5,547,835 describe methods
5 of sequencing nucleic acids using mass spectrometry. US Patent Nos. 5,622,824, 5,872,003 and 5,691,141 describe methods, systems and kits for exonuclease-mediated mass spectrometric sequencing.

Thus, there is a need for a method for bioagent detection and identification which is both specific and rapid, and in which no nucleic acid sequencing is required.
10 The present invention addresses this need.

SUMMARY OF THE INVENTION

The present invention is directed to method of identifying an unknown bioagent using a database, such as a database stored on, for example, a local computer
15 or perhaps a database accessible over a network or on the internet. This database of molecular masses of known bioagents provides a standard of comparison for determining both identity and geographic origin of the unknown bioagent. The nucleic acid from said bioagent is first contacted with at least one pair of oligonucleotide primers which hybridize to sequences of said nucleic acid that flank a variable nucleic
20 acid sequence of the bioagent. Using PCR technology an amplification product of this variable nucleic acid sequence is made. After standard isolation, the molecular mass of this amplification product is determined using known mass-spec techniques. This molecular mass is compared to the molecular mass of known bioagents within the database, for identifying the unknown bioagent.

25 This invention is also directed to databases having cell-data positional significance comprising at least a first table that includes a plurality of data-containing cells. The table is organized into at least a first row and a second row, each row having columns which are aligned relative to each other so that inter-row conserved regions are aligned. This alignment facilitates the analysis of regions, which are highly
30 conserved between species. This alignment further provides insight into secondary polymer structure by this alignment. Although this invention is directed to a database where each row describes any polymer, in a preferred embodiment, the polymer is an RNA. Other alignments that operate in the same manner are also contemplated.

Another embodiment of this invention is a method for reconciling the content of two databases such that the content of each is a mirror of the other.

Another embodiment is directed to determining the geographic origin of a bioagent using a database of molecular masses of known bioagents comprising
5 contacting a nucleic acid from the selected bioagent with at least one pair of oligonucleotide primers which hybridize to sequences of the nucleic acid, where the sequences flank a variable nucleic acid sequence of the bioagent. This hybridized region is isolated and amplified through standard PCR techniques known in the art. The molecular mass is determined of this amplified product through any technique
10 known in the art such as, Mass-spectrometry for example. This molecular mass is compared to the molecular masses stored in the database of known bioagents thereby determining a group of probabilistically reasonable geographic origins for the selected bioagent.

15 BRIEF DESCRIPTION OF THE DRAWINGS

Figures 1A-1I are representative consensus diagrams that show examples of conserved regions from 16S rRNA (Fig. 1A-1B), 23S rRNA (3'-half, Fig. 1C-1D; 5'-half, Fig. 1E-F), 23S rRNA Domain I (Fig. 1G), 23S rRNA Domain IV (Fig. 1H) and 16S rRNA Domain III (Fig. 1I) which are suitable for use in the present invention.
20 Lines with arrows are examples of regions to which intelligent primer pairs for PCR are designed. The label for each primer pair represents the starting and ending base number of the amplified region on the consensus diagram. Bases in capital letters are greater than 95% conserved; bases in lower case letters are 90-95% conserved, filled circles are 80-90% conserved; and open circles are less than 80% conserved. The label
25 for each primer pair represents the starting and ending base number of the amplified region on the consensus diagram.

Figure 2 shows a typical primer amplified region from the 16S rRNA Domain III shown in Figure 1C.

Figure 3 is a schematic diagram showing conserved regions in RNase P.
30 Bases in capital letters are greater than 90% conserved; bases in lower case letters are 80-90% conserved; filled circles designate bases which are 70-80% conserved; and open circles designate bases that are less than 70% conserved.

Figure 4 is a schematic diagram of base composition signature determination using nucleotide analog "tags" to determine base composition signatures.

Figure 5 shows the deconvoluted mass spectra of a *Bacillus anthracis* region with and without the mass tag phosphorothioate A (A*). The two spectra differ in that the measured molecular weight of the mass tag-containing sequence is greater than the unmodified sequence.

5 Figure 6 shows base composition signature (BCS) spectra from PCR products from *Staphylococcus aureus* (*S. aureus* 16S_1337F) and *Bacillus anthracis* (*B. anthr.* 16S_1337F), amplified using the same primers. The two strands differ by only two (AT-->CG) substitutions and are clearly distinguished on the basis of their BCS.

10 Figure 7 shows that a single difference between two sequences (A14 in *B. anthracis* vs. A15 in *B. cereus*) can be easily detected using ESI-TOF mass spectrometry.

Figure 8 is an ESI-TOF of *Bacillus anthracis* spore coat protein sspE 56mer plus calibrant. The signals unambiguously identify *B. anthracis* versus other *Bacillus* species.

15 Figure 9 is an ESI-TOF of a *B. anthracis* synthetic 16S_1228 duplex (reverse and forward strands). The technique easily distinguishes between the forward and reverse strands.

Figure 10 is an ESI-FTICR-MS of a synthetic *B. anthracis* 16S_1337 46 base pair duplex.

20 Figure 11 is an ESI-TOF-MS of a 56mer oligonucleotide (3 scans) from the *B. anthracis* saspB gene with an internal mass standard. The internal mass standards are designated by asterisks.

25 Figure 12 is an ESI-TOF-MS of an internal standard with 5 mM TBA-TFA buffer showing that charge stripping with tributylammonium trifluoroacetate reduces the most abundant charge state from [M-8H+]⁸⁻ to [M-3H+]³⁻.

Figure 13 is a portion of a secondary structure defining database according to one embodiment of the present invention, where two examples of selected sequences are displayed graphically thereunder.

30 Figure 14 is a three dimensional graph demonstrating the grouping of sample molecular weight according to species.

Figure 15 is a three dimensional graph demonstrating the grouping of sample molecular weights according to species of virus and mammal infected.

Figure 16 is a three dimensional graph demonstrating the grouping of sample molecular weights according to species of virus, and animal-origin of infectious agent.

Figure 17 is a figure depicting how the triangulation method of the present invention provides for the identification of an unknown bioagent without prior knowledge of the unknown agent. The use of different primer sets to distinguish and identify the unknown is also depicted as primer sets I, II and III within this figure. A three dimensional graph depicts all of bioagent space (170), including the unknown bioagent, which after use of primer set I (171) according to a method according to the present invention further differentiates and classifies bioagents according to major classifications (176) which, upon further analysis using primer set II (172) differentiates the unknown agent (177) from other, known agents (173) and finally, the use of a third primer set (175) further specifies subgroups within the family of the unknown (174).

DESCRIPTION OF PREFERRED EMBODIMENTS

The present invention provides a combination of a non-PCR biomass detection mode, preferably high-resolution MS, with PCR-based BCS technology using “intelligent primers” which hybridize to conserved sequence regions of nucleic acids derived from a bioagent and which bracket variable sequence regions that uniquely identify the bioagent. The high-resolution MS technique is used to determine the molecular mass and base composition signature (BCS) of the amplified sequence region. This unique “base composition signature” (BCS) is then input to a maximum-likelihood detection algorithm for matching against a database of base composition signatures in the same amplified region. The present method combines PCR-based amplification technology (which provides specificity) and a molecular mass detection mode (which provides speed and does not require nucleic acid sequencing of the amplified target sequence) for bioagent detection and identification.

The present methods allow extremely rapid and accurate detection and identification of bioagents compared to existing methods. Furthermore, this rapid detection and identification is possible even when sample material is impure. Thus, the method is useful in a wide variety of fields, including, but not limited to, environmental testing (e.g., detection and discrimination of pathogenic vs. non-pathogenic bacteria in water or other samples), germ warfare (allowing immediate identification of the bioagent and appropriate treatment), pharmacogenetic analysis and medical diagnosis (including cancer diagnosis based on mutations and polymorphisms; drug resistance and susceptibility testing; screening for and/or

diagnosis of genetic diseases and conditions; and diagnosis of infectious diseases and conditions). The methods leverage ongoing biomedical research in virulence, pathogenicity, drug resistance and genome sequencing into a method which provides greatly improved sensitivity, specificity and reliability compared to existing methods, with lower rates of false positives.

The present methods can be used, for example, to detect and classify any biological agent, including bacteria, viruses, fungi and toxins. As one example, where the agent is a biological threat, the information obtained is used to determine practical information needed for countermeasures, including toxin genes, pathogenicity islands and antibiotic resistance genes. In addition, the methods can be used to identify natural or deliberate engineering events including chromosome fragment swapping, molecular breeding (gene shuffling) and emerging infectious diseases.

Bacteria have a common set of absolutely required genes. About 250 genes are present in all bacterial species (*Proc. Natl. Acad. Sci. U.S.A.*, **1996**, 93, 10268; *Science*, **1995**, 270, 397), including tiny genomes like *Mycoplasma*, *Ureaplasma* and *Rickettsia*. These genes encode proteins involved in translation, replication, recombination and repair, transcription, nucleotide metabolism, amino acid metabolism, lipid metabolism, energy generation, uptake, secretion and the like. Examples of these proteins are DNA polymerase III beta, elongation factor TU, heat shock protein groEL, RNA polymerase beta, phosphoglycerate kinase, NADH dehydrogenase, DNA ligase, DNA topoisomerase and elongation factor G. Operons can also be targeted using the present method. One example of an operon is the bfp operon from enteropathogenic *E. coli*. Multiple core chromosomal genes can be used to classify bacteria at a genus or genus species level to determine if an organism has threat potential. The methods can also be used to detect pathogenicity markers (plasmid or chromosomal) and antibiotic resistance genes to confirm the threat potential of an organism and to direct countermeasures.

A theoretically ideal bioagent detector would identify, quantify, and report the complete nucleic acid sequence of every bioagent that reached the sensor. The complete sequence of the nucleic acid component of a pathogen would provide all relevant information about the threat, including its identity and the presence of drug-resistance or pathogenicity markers. This ideal has not yet been achieved. However, the present invention provides a straightforward strategy for obtaining information with the same practical value using base composition signatures (BCS). While the

base composition of a gene fragment is not as information-rich as the sequence itself, there is no need to analyze the complete sequence of the gene if the short analyte sequence fragment is properly chosen. A database of reference sequences can be prepared in which each sequence is indexed to a unique base composition signature, so that the presence of the sequence can be inferred with accuracy from the presence of the signature. The advantage of base composition signatures is that they can be quantitatively measured in a massively parallel fashion using multiplex PCR (PCR in which two or more primer pairs amplify target sequences simultaneously) and mass spectrometry. These multiple primer amplified regions uniquely identify most threat and ubiquitous background bacteria and viruses. In addition, cluster-specific primer pairs distinguish important local clusters (e.g., anthracis group).

In the context of this invention, a "bioagent" is any organism, living or dead, or a nucleic acid derived from such an organism. Examples of bioagents include, but are not limited to, cells (including, but not limited to, human clinical samples, bacterial cells and other pathogens), viruses, toxin genes and bioregulating compounds. Samples may be alive or dead or in a vegetative state (for example, vegetative bacteria or spores) and may be encapsulated or bioengineered.

As used herein, a "base composition signature" (BCS) is the exact base composition from selected fragments of nucleic acid sequences that uniquely identifies the target gene and source organism. BCS can be thought of as unique indexes of specific genes.

As used herein, "intelligent primers" are primers which bind to sequence regions which flank an intervening variable region. In a preferred embodiment, these sequence regions which flank the variable region are highly conserved among different species of bioagent. For example, the sequence regions may be highly conserved among all *Bacillus* species. By the term "highly conserved," it is meant that the sequence regions exhibit between about 80-100%, more preferably between about 90-100% and most preferably between about 95-100% identity. Examples of intelligent primers which amplify regions of the 16S and 23S rRNA are shown in Figures 1A-II. A typical primer amplified region in 16S rRNA is shown in Figure 2. The arrows represent primers which bind to highly conserved regions which flank a variable region in 16S rRNA domain III. The amplified region is the stem-loop structure under "1100-1188."

One main advantage of the detection methods of the present invention is that the primers need not be specific for a particular bacterial species, or even genus, such as *Bacillus* or *Streptomyces*. Instead, the primers recognize highly conserved regions across hundreds of bacterial species including, but not limited to, the species described
5 herein. Thus, the same primer pair can be used to identify any desired bacterium because it will bind to the conserved regions which flank a variable region specific to a single species, or common to several bacterial species, allowing nucleic acid amplification of the intervening sequence and determination of its molecular weight and base composition. For example, the 16S_971-1062, 16S_1228-1310 and
10 16S_1100-1188 regions are 98-99% conserved in about 900 species of bacteria (16S=16S rRNA, numbers indicate nucleotide position). In one embodiment of the present invention, primers used in the present method bind to one or more of these regions or portions thereof.

The present invention provides a combination of a non-PCR biomass
15 detection mode, preferably high-resolution MS, with nucleic acid amplification-based BCS technology using "intelligent primers" which hybridize to conserved regions and which bracket variable regions that uniquely identify the bioagent(s). Although the use of PCR is preferred, other nucleic acid amplification techniques may also be used, including ligase chain reaction (LCR) and strand displacement amplification (SDA).
20 The high-resolution MS technique allows separation of bioagent spectral lines from background spectral lines in highly cluttered environments. The resolved spectral lines are then translated to BCS which are input to a maximum-likelihood detection algorithm matched against spectra for one or more known BCS. Preferably, the bioagent BCS spectrum is matched against one or more databases of BCS from vast
25 numbers of bioagents. Preferably, the matching is done using a maximum-likelihood detection algorithm.

In one embodiment, base composition signatures are quantitatively measured in a massively parallel fashion using the polymerase chain reaction (PCR), preferably multiplex PCR, and mass spectrometric (MS) methods. Sufficient quantities of nucleic
30 acids should be present for detection of bioagents by MS. A wide variety of techniques for preparing large amounts of purified nucleic acids or fragments thereof are well known to those of skill in the art. PCR requires one or more pairs of oligonucleotide primers which bind to regions which flank the target sequence(s) to be amplified. These primers prime synthesis of a different strand of DNA, with synthesis

occurring in the direction of one primer towards the other primer. The primers, DNA to be amplified, a thermostable DNA polymerase (e.g. *Taq* polymerase), the four deoxynucleotide triphosphates, and a buffer are combined to initiate DNA synthesis. The solution is denatured by heating, then cooled to allow annealing of newly added
5 primer, followed by another round of DNA synthesis. This process is typically repeated for about 30 cycles, resulting in amplification of the target sequence.

The “intelligent primers” define the target sequence region to be amplified and analyzed. In one embodiment, the target sequence is a ribosomal RNA (rRNA) gene sequence. With the complete sequences of many of the smallest microbial
10 genomes now available, it is possible to identify a set of genes that defines “minimal life” and identify composition signatures that uniquely identify each gene and organism. Genes that encode core life functions such as DNA replication, transcription, ribosome structure, translation, and transport are distributed broadly in the bacterial genome and are preferred regions for BCS analysis. Ribosomal RNA
15 (rRNA) genes comprise regions that provide useful base composition signatures. Like many genes involved in core life functions, rRNA genes contain sequences that are extraordinarily conserved across bacterial domains interspersed with regions of high variability that are more specific to each species. The variable regions can be utilized to build a database of base composition signatures. The strategy involves creating a
20 structure-based alignment of sequences of the small (16S) and the large (23S) subunits of the rRNA genes. For example, there are currently over 13,000 sequences in the ribosomal RNA database that has been created and maintained by Robin Gutell, University of Texas at Austin, and is publicly available on the Institute for Cellular and Molecular Biology web page on the world wide web of the Internet at, for
25 example, “rna.icmb.utexas.edu/.” There is also a publicly available rRNA database created and maintained by the University of Antwerp, Belgium on the world wide web of the Internet at, for example, “rna.uia.ac.be.”

These databases have been analyzed to determine regions that are useful as base composition signatures. The characteristics of such regions include: a) between
30 about 80 and 100%, preferably > about 95% identity among species of the particular bioagent of interest, of upstream and downstream nucleotide sequences which serve as sequence amplification primer sites; b) an intervening variable region which exhibits no greater than about 5% identity among species; and c) a separation of between about 30 and 1000 nucleotides, preferably no more than about 50-250 nucleotides, and more

preferably no more than about 60-100 nucleotides, between the conserved regions.

Due to their overall conservation, the flanking rRNA primer sequences serve as good “universal” primer binding sites to amplify the region of interest for most, if not all, bacterial species. The intervening region between the sets of primers varies in
5 length and/or composition, and thus provides a unique base composition signature.

It is advantageous to design the “intelligent primers” to be as universal as possible to minimize the number of primers which need to be synthesized, and to allow detection of multiple species using a single pair of primers. These primer pairs can be used to amplify variable regions in these species. Because any variation (due to
10 codon wobble in the 3rd position) in these conserved regions among species is likely to occur in the third position of a DNA triplet, oligonucleotide primers can be designed such that the nucleotide corresponding to this position is a base which can bind to more than one nucleotide, referred to herein as a “universal base.” For example, under this “wobble” pairing, inosine (I) binds to U, C or A; guanine (G) binds to U or C, and
15 uridine (U) binds to U or C. Other examples of universal bases include nitroindoles such as 5-nitroindole or 3-nitropyrrole (Loakes *et al.*, *Nucleosides and Nucleotides*, **1995**, 14, 1001-1003), the degenerate nucleotides dP or dK (Hill *et al.*), an acyclic nucleoside analog containing 5-nitroindazole (Van Aerschot *et al.*, *Nucleosides and Nucleotides*, **1995**, 14, 1053-1056) or the purine analog 1-(2-deoxy- β -D-
20 ribofuranosyl)-imidazole-4-carboxamide (Sala *et al.*, *Nucl. Acids Res.*, **1996**, 24, 3302-3306).

In another embodiment of the invention, to compensate for the somewhat weaker binding by the “wobble” base, the oligonucleotide primers are designed such that the first and second positions of each triplet are occupied by nucleotide analogs
25 which bind with greater affinity than the unmodified nucleotide. Examples of these analogs include, but are not limited to, 2,6-diaminopurine which binds to thymine, propyne T which binds to adenine and propyne C and phenoxazines, including G-clamp, which binds to G. Propynes are described in U.S. Patent Nos. 5,645,985, 5,830,653 and 5,484,908, each of which is incorporated herein by reference in its
30 entirety. Phenoxazines are described in U.S. Patent Nos. 5,502,177, 5,763,588, and 6,005,096, each of which is incorporated herein by reference in its entirety. G-clamps are described in U.S. Patent Nos. 6,007,992 and 6,028,183, each of which is incorporated herein by reference in its entirety.

Bacterial biological warfare agents capable of being detected by the present

methods include, but are not limited to, *Bacillus anthracis* (anthrax), *Yersinia pestis* (pneumonic plague), *Francisella tularensis* (tularemia), *Brucella suis*, *Brucella abortus*, *Brucella melitensis* (undulant fever), *Burkholderia mallei* (glanders), *Burkholderia pseudomallei* (melioidosis), *Salmonella typhi* (typhoid fever), *Rickettsia typhi* (epidemic typhus), *Rickettsia prowasekii* (endemic typhus) and *Coxiella burnetii* (Q fever), *Rhodobacter capsulatus*, *Chlamydia pneumoniae*, *Escherichia coli*, *Shigella dysenteriae*, *Shigella flexneri*, *Bacillus cereus*, *Clostridium botulinum*, *Coxiella burnetii*, *Pseudomonas aeruginosa*, *Legionella pneumophila*, and *Vibrio cholerae*.

Besides 16S and 23S rRNA, other target regions suitable for use in the present invention for detection of bacteria include, but are not limited to, 5S rRNA and RNase P (Figure 3).

Biological warfare fungus biowarfare agents include, but are not limited to, *coccidioides immitis* (Coccidioidomycosis).

Biological warfare toxin genes capable of being detected by the methods of the present invention include, but are not limited to, botulism, T-2 mycotoxins, ricin, staph enterotoxin B, shigatoxin, abrin, aflatoxin, *Clostridium perfringens* epsilon toxin, conotoxins, diacetoxyscirpenol, tetrodotoxin and saxitoxin.

Biological warfare viral threat agents are mostly RNA viruses (positive-strand and negative-strand), with the exception of smallpox. Every RNA virus is a family of related viruses (quasispecies). These viruses mutate rapidly and the potential for engineered strains (natural or deliberate) is very high. RNA viruses cluster into families that have conserved RNA structural domains on the viral genome (e.g., virion components, accessory proteins) and conserved housekeeping genes that encode core viral proteins including, for single strand positive strand RNA viruses, RNA-dependent RNA polymerase, double stranded RNA helicase, chymotrypsin-like and papain-like proteases and methyltransferases.

Examples of (-)-strand RNA viruses include, but are not limited to, arenaviruses (e.g., sabia virus, lassa fever, Machupo, Argentine hemorrhagic fever, flexal virus), bunyaviruses (e.g., hantavirus, nairovirus, phlebovirus, hantaan virus, Congo-crimean hemorrhagic fever, rift valley fever), and mononegavirales (e.g., filovirus, paramyxovirus, ebola virus, Marburg, equine morbillivirus).

Examples of (+)-strand RNA viruses include, but are not limited to, picornaviruses (e.g., coxsackievirus, echovirus, human coxsackievirus A, human

echovirus, human enterovirus, human poliovirus, hepatitis A virus, human parechovirus, human rhinovirus), astroviruses (e.g., human astrovirus), calciviruses (e.g., chiba virus, chitta virus, human calcivirus, norwalk virus), nidovirales (e.g., human coronavirus, human torovirus), flaviviruses (e.g., dengue virus 1-4, Japanese encephalitis virus, Kyanasur forest disease virus, Murray Valley encephalitis virus, Rocio virus, St. Louis encephalitis virus, West Nile virus, yellow fever virus, hepatitis c virus) and togaviruses (e.g., Chikugunya virus, Eastern equine encephalitis virus, Mayaro virus, O'nyong-nyong virus, Ross River virus, Venezuelan equine encephalitis virus, Rubella virus, hepatitis E virus). The hepatitis C virus has a 5'-untranslated region of 340 nucleotides, an open reading frame encoding 9 proteins having 3010 amino acids and a 3'-untranslated region of 240 nucleotides. The 5'-UTR and 3'-UTR are 99% conserved in hepatitis C viruses.

In one embodiment, the target gene is an RNA-dependent RNA polymerase or a helicase encoded by (+)-strand RNA viruses, or RNA polymerase from a (-)-strand RNA virus. (+)-strand RNA viruses are double stranded RNA and replicate by RNA-directed RNA synthesis using RNA-dependent RNA polymerase and the positive strand as a template. Helicase unwinds the RNA duplex to allow replication of the single stranded RNA. These viruses include viruses from the family picornaviridae (e.g., poliovirus, coxsackievirus, echovirus), togaviridae (e.g., alphavirus, flavivirus, rubivirus), arenaviridae (e.g., lymphocytic choriomeningitis virus, lassa fever virus), cononaviridae (e.g., human respiratory virus) and Hepatitis A virus. The genes encoding these proteins comprise variable and highly conserved regions which flank the variable regions.

In another embodiment, the detection scheme for the PCR products generated from the bioagent(s) incorporates at least three features. First, the technique simultaneously detects and differentiates multiple (generally about 6-10) PCR products. Second, the technique provides a BCS that uniquely identifies the bioagent from the possible primer sites. Finally, the detection technique is rapid, allowing multiple PCR reactions to be run in parallel.

In one embodiment, the method can be used to detect the presence of antibiotic resistance and/or toxin genes in a bacterial species. For example, *Bacillus anthracis* comprising a tetracycline resistance plasmid and plasmids encoding one or both anthracis toxins (px01 and/or px02) can be detected by using antibiotic resistance primer sets and toxin gene primer sets. If the *B. anthracis* is positive for tetracycline

resistance, then a different antibiotic, for example quinalone, is used.

Mass spectrometry (MS)-based detection of PCR products provides all of these features with additional advantages. MS is intrinsically a parallel detection scheme without the need for radioactive or fluorescent labels, since every amplification product with a unique base composition is identified by its molecular mass. The current state of the art in mass spectrometry is such that less than femtomole quantities of material can be readily analyzed to afford information about the molecular contents of the sample. An accurate assessment of the molecular mass of the material can be quickly obtained, irrespective of whether the molecular weight of the sample is several hundred, or in excess of one hundred thousand atomic mass units (amu) or Daltons. Intact molecular ions can be generated from amplification products using one of a variety of ionization techniques to convert the sample to gas phase. These ionization methods include, but are not limited to, electrospray ionization (ESI), matrix-assisted laser desorption ionization (MALDI) and fast atom bombardment (FAB). For example, MALDI of nucleic acids, along with examples of matrices for use in MALDI of nucleic acids, are described in WO 98/54751 (Genetrace, Inc.).

Upon ionization, several peaks are observed from one sample due to the formation of ions with different charges. Averaging the multiple readings of molecular mass obtained from a single mass spectrum affords an estimate of molecular mass of the bioagent. Electrospray ionization mass spectrometry (ESI-MS) is particularly useful for very high molecular weight polymers such as proteins and nucleic acids having molecular weights greater than 10 kDa, since it yields a distribution of multiply-charged molecules of the sample without causing a significant amount of fragmentation.

The mass detectors used in the methods of the present invention include, but are not limited to, Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR-MS), ion trap, quadrupole, magnetic sector, time of flight (TOF), Q-TOF, and triple quadrupole.

In general, the mass spectrometric techniques which can be used in the present invention include, but are not limited to, tandem mass spectrometry, infrared multiphoton dissociation and pyrolytic gas chromatography mass spectrometry (PGC-MS). In one embodiment of the invention, the bioagent detection system operates continually in bioagent detection mode using pyrolytic GC-MS without PCR for rapid

detection of increases in biomass (for example, increases in fecal contamination of drinking water or of germ warfare agents). To achieve minimal latency, a continuous sample stream flows directly into the PGC-MS combustion chamber. When an increase in biomass is detected, a PCR process is automatically initiated. Bioagent
5 presence produces elevated levels of large molecular fragments from, for example, about 100-7,000 Da which are observed in the PGC-MS spectrum. The observed mass spectrum is compared to a threshold level and when levels of biomass are determined to exceed a predetermined threshold, the bioagent classification process described hereinabove (combining PCR and MS, preferably FT-ICR MS) is initiated.
10 Optionally, alarms or other processes (halting ventilation flow, physical isolation) are also initiated by this detected biomass level.

The accurate measurement of molecular mass for large DNAs is limited by the adduction of cations from the PCR reaction to each strand, resolution of the isotopic peaks from natural abundance ^{13}C and ^{15}N isotopes, and assignment of the
15 charge state for any ion. The cations are removed by in-line dialysis using a flow-through chip that brings the solution containing the PCR products into contact with a solution containing ammonium acetate in the presence of an electric field gradient orthogonal to the flow. The latter two problems are addressed by operating with a resolving power of >100,000 and by incorporating isotopically depleted nucleotide
20 triphosphates into the DNA. The resolving power of the instrument is also a consideration. At a resolving power of 10,000, the modeled signal from the $[\text{M}-14\text{H}]^{14+}$ charge state of an 84mer PCR product is poorly characterized and assignment of the charge state or exact mass is impossible. At a resolving power of 33,000, the peaks from the individual isotopic components are visible. At a resolving
25 power of 100,000, the isotopic peaks are resolved to the baseline and assignment of the charge state for the ion is straightforward. The $^{13}\text{C}, ^{15}\text{N}$ -depleted triphosphates are obtained, for example, by growing microorganisms on depleted media and harvesting the nucleotides (Batey *et al.*, *Nucl. Acids Res.*, **1992**, 20, 4515-4523).

While mass measurements of intact nucleic acid regions are believed to be
30 adequate to determine most bioagents, tandem mass spectrometry (MS^n) techniques may provide more definitive information pertaining to molecular identity or sequence. Tandem MS involves the coupled use of two or more stages of mass analysis where both the separation and detection steps are based on mass spectrometry. The first stage is used to select an ion or component of a sample from which further structural

information is to be obtained. The selected ion is then fragmented using, e.g., blackbody irradiation, infrared multiphoton dissociation, or collisional activation. For example, ions generated by electrospray ionization (ESI) can be fragmented using IR multiphoton dissociation. This activation leads to dissociation of glycosidic bonds and the phosphate backbone, producing two series of fragment ions, called the *w*-series (having an intact 3' terminus and a 5' phosphate following internal cleavage) and the α -Base series (having an intact 5' terminus and a 3' furan).

The second stage of mass analysis is then used to detect and measure the mass of these resulting fragments of product ions. Such ion selection followed by fragmentation routines can be performed multiple times so as to essentially completely dissect the molecular sequence of a sample.

If there are two or more targets of similar base composition or mass, or if a single amplification reaction results in a product which has the same mass as two or more bioagent reference standards, they can be distinguished by using mass-modifying "tags." In this embodiment of the invention, a nucleotide analog or "tag" is incorporated during amplification (e.g., a 5-(trifluoromethyl) deoxythymidine triphosphate) which has a different molecular weight than the unmodified base so as to improve distinction of masses. Such tags are described in, for example, PCT WO97/33000, which is incorporated herein by reference in its entirety. This further limits the number of possible base compositions consistent with any mass. For example, 5-(trifluoromethyl)deoxythymidine triphosphate can be used in place of dTTP in a separate nucleic acid amplification reaction. Measurement of the mass shift between a conventional amplification product and the tagged product is used to quantitate the number of thymidine nucleotides in each of the single strands. Because the strands are complementary, the number of adenosine nucleotides in each strand is also determined.

In another amplification reaction, the number of G and C residues in each strand is determined using, for example, the cytidine analog 5-methylcytosine (5-meC) or propyne C. The combination of the A/T reaction and G/C reaction, followed by molecular weight determination, provides a unique base composition. This method is summarized in Figure 4 and Table 1.

Table 1

| Mass tag | Double strand sequence | Single strand sequence | Total mass this strand | Base info this strand | Base info other strand | Total base comp. Top strand | Total base comp. Bottom strand |
|-----------------------|-------------------------------------|------------------------|------------------------|-----------------------|------------------------|-----------------------------|--------------------------------|
| T*.mass (T*-T) = x | T*ACGT*A CGT* AT*GCAT* GCA | T*ACGT*A CGT* | 3x | 3T | 3A | 3T 2A 2C 2G | 3A 2T 2G 2C |
| | | AT*GCAT* GCA | 2x | 2T | 2A | | |
| C*.mass (C*-C) = y | TAC*GTAC *GT ATGC*ATG C*A | TAC*GTAC *GT | 2x | 2C | 2G | | |
| | | ATGC*ATG C*A | 2x | 2C | 2G | | |

The mass tag phosphorothioate A (A*) was used to distinguish a *Bacillus anthracis* cluster. The *B. anthracis* (A₁₄G₉C₁₄T₉) had an average MW of 14072.26, and the *B. anthracis* (A₁A*₁₃G₉C₁₄T₉) had an average molecular weight of 14281.11 and the phosphorothioate A had an average molecular weight of +16.06 as determined by ESI-TOF MS. The deconvoluted spectra are shown in Figure 5.

In another example, assume the measured molecular masses of each strand are 30,000.115Da and 31,000.115 Da respectively, and the measured number of dT and dA residues are (30,28) and (28,30). If the molecular mass is accurate to 100 ppm, there are 7 possible combinations of dG+dC possible for each strand. However, if the measured molecular mass is accurate to 10 ppm, there are only 2 combinations of dG+dC, and at 1 ppm accuracy there is only one possible base composition for each strand.

Signals from the mass spectrometer may be input to a maximum-likelihood detection and classification algorithm such as is widely used in radar signal processing. The detection processing uses matched filtering of BCS observed in mass-basecount space and allows for detection and subtraction of signatures from known, harmless organisms, and for detection of unknown bioagent threats. Comparison of newly observed bioagents to known bioagents is also possible, for estimation of threat level, by comparing their BCS to those of known organisms and to known forms of pathogenicity enhancement, such as insertion of antibiotic resistance genes or toxin genes.

Processing may end with a Bayesian classifier using log likelihood ratios developed from the observed signals and average background levels. The program emphasizes performance predictions culminating in probability-of-detection versus probability-of-false-alarm plots for conditions involving complex backgrounds of naturally occurring organisms and environmental contaminants. Matched filters consist of a priori expectations of signal values given the set of primers used for each of the bioagents. A genomic sequence database (e.g. GenBank) is used to define the mass basecount matched filters. The database contains known threat agents and benign background organisms. The latter is used to estimate and subtract the signature produced by the background organisms. A maximum likelihood detection of known background organisms is implemented using matched filters and a running-sum estimate of the noise covariance. Background signal strengths are estimated and used along with the matched filters to form signatures which are then subtracted. the maximum likelihood process is applied to this "cleaned up" data in a similar manner employing matched filters for the organisms and a running-sum estimate of the noise-covariance for the cleaned up data.

In one embodiment, a strategy to "triangulate" each organism by measuring signals from multiple core genes is used to reduce false negative and false positive signals, and enable reconstruction of the origin or hybrid or otherwise engineered bioagents. After identification of multiple core genes, alignments are created from nucleic acid sequence databases. The alignments are then analyzed for regions of conservation and variation, and potential primer binding sites flanking variable regions are identified. Next, amplification target regions for signature analysis are selected which distinguishes organisms based on specific genomic differences (i.e., base composition). For example, detection of signatures for the three part toxin genes

typical of *B. anthracis* (Bowen *et al.*, *J. Appl. Microbiol.*, **1999**, 87, 270-278) in the absence of the expected signatures from the *B. anthracis* genome would suggest a genetic engineering event.

The present method can also be used to detect single nucleotide polymorphisms (SNPs), or multiple nucleotide polymorphisms, rapidly and accurately. A SNP is defined as a single base pair site in the genome that is different from one individual to another. The difference can be expressed either as a deletion, an insertion or a substitution, and is frequently linked to a disease state. Because they occur every 100-1000 base pairs, SNPs are the most frequently bound type of genetic marker in the human genome.

For example, sickle cell anemia results from an A-T transition, which encodes a valine rather than a glutamic acid residue. Oligonucleotide primers may be designed such that they bind to sequences which flank a SNP site, followed by nucleotide amplification and mass determination of the amplified product. Because the molecular masses of the resulting product from an individual who does not have sickle cell anemia is different from that of the product from an individual who has the disease, the method can be used to distinguish the two individuals. Thus, the method can be used to detect any known SNP in an individual and thus diagnose or determine increased susceptibility to a disease or condition.

In one embodiment, blood is drawn from an individual and peripheral blood mononuclear cells (PBMC) are isolated and simultaneously tested, preferably in a high-throughput screening method, for one or more SNPs using appropriate primers based on the known sequences which flank the SNP region. The National Center for Biotechnology Information maintains a publicly available database of SNPs on the world wide web of the Internet at, for example, "ncbi.nlm.nih.gov/SNP/."

The method of the present invention can also be used for blood typing. The gene encoding A, B or O blood type can differ by four single nucleotide polymorphisms. If the gene contains the sequence CGTGGTGACCCTT (SEQ ID NO:1), antigen A results. If the gene contains the sequence CGTCGTCACCGCTA (SEQ ID NO:2) antigen B results. If the gene contains the sequence CGTGGT-ACCCCTT (SEQ ID NO:3), blood group O results ("-" indicates a deletion). These sequences can be distinguished by designing a single primer pair which flanks these regions, followed by amplification and mass determination.

While the present invention has been described with specificity in accordance with certain of its preferred embodiments, the following examples serve only to illustrate the invention and are not intended to limit the same.

5 **EXAMPLES**

Example 1: Nucleic Acid Isolation and PCR

In one embodiment, nucleic acid is isolated from the organisms and amplified by PCR using standard methods prior to BCS determination by mass spectrometry. Nucleic acid is isolated, for example, by detergent lysis of bacterial cells, centrifugation and ethanol precipitation. Nucleic acid isolation methods are described in, for example, *Current Protocols in Molecular Biology* (Ausubel et al.) and *Molecular Cloning; A Laboratory Manual* (Sambrook et al.). The nucleic acid is then amplified using standard methodology, such as PCR, with primers which bind to conserved regions of the nucleic acid which contain an intervening variable sequence as described below.

Example 2: Mass spectrometry

FTICR Instrumentation: The FTICR instrument is based on a 7 tesla actively shielded superconducting magnet and modified Bruker Daltonics Apex II 70e ion optics and vacuum chamber. The spectrometer is interfaced to a LEAP PAL autosampler and a custom fluidics control system for high throughput screening applications. Samples are analyzed directly from 96-well or 384-well microtiter plates at a rate of about 1 sample/minute. The Bruker data-acquisition platform is supplemented with a lab-built ancillary NT datastation which controls the autosampler and contains an arbitrary waveform generator capable of generating complex rf-excite waveforms (frequency sweeps, filtered noise, stored waveform inverse Fourier transform (SWIFT), etc.) for sophisticated tandem MS experiments. For oligonucleotides in the 20-30-mer regime typical performance characteristics include mass resolving power in excess of 100,000 (FWHM), low ppm mass measurement errors, and an operable m/z range between 50 and 5000 m/z .

Modified ESI Source: In sample-limited analyses, analyte solutions are delivered at 150 nL/minute to a 30 mm i.d. fused-silica ESI emitter mounted on a 3-D micromanipulator. The ESI ion optics consist of a heated metal capillary, an rf-only hexapole, a skimmer cone, and an auxiliary gate electrode. The 6.2 cm rf-only

hexapole is comprised of 1 mm diameter rods and is operated at a voltage of 380 Vpp at a frequency of 5 MHz. A lab-built electro-mechanical shutter can be employed to prevent the electrospray plume from entering the inlet capillary unless triggered to the “open” position via a TTL pulse from the data station. When in the “closed” position, a stable electrospray plume is maintained between the ESI emitter and the face of the shutter. The back face of the shutter arm contains an elastomeric seal which can be positioned to form a vacuum seal with the inlet capillary. When the seal is removed, a 1 mm gap between the shutter blade and the capillary inlet allows constant pressure in the external ion reservoir regardless of whether the shutter is in the open or closed position. When the shutter is triggered, a “time slice” of ions is allowed to enter the inlet capillary and is subsequently accumulated in the external ion reservoir. The rapid response time of the ion shutter (<25 ms) provides reproducible, user defined intervals during which ions can be injected into and accumulated in the external ion reservoir.

Apparatus for Infrared Multiphoton Dissociation: A 25 watt CW CO₂ laser operating at 10.6 μm has been interfaced to the spectrometer to enable infrared multiphoton dissociation (IRMPD) for oligonucleotide sequencing and other tandem MS applications. An aluminum optical bench is positioned approximately 1.5 m from the actively shielded superconducting magnet such that the laser beam is aligned with the central axis of the magnet. Using standard IR-compatible mirrors and kinematic mirror mounts, the unfocused 3 mm laser beam is aligned to traverse directly through the 3.5 mm holes in the trapping electrodes of the FTICR trapped ion cell and longitudinally traverse the hexapole region of the external ion guide finally impinging on the skimmer cone. This scheme allows IRMPD to be conducted in an *m/z* selective manner in the trapped ion cell (e.g. following a SWIFT isolation of the species of interest), or in a broadband mode in the high pressure region of the external ion reservoir where collisions with neutral molecules stabilize IRMPD-generated metastable fragment ions resulting in increased fragment ion yield and sequence coverage.

30 **Example 3 :Identification of Bioagents**

Table 2 shows a small cross section of a database of calculated molecular masses for over 9 primer sets and approximately 30 organisms. The primer sets were derived from rRNA alignment. Examples of regions from rRNA consensus alignments are shown in Figures 1A-1C. Lines with arrows are examples of regions to which

intelligent primer pairs for PCR are designed. The primer pairs are >95% conserved in the bacterial sequence database (currently over 10,000 organisms). The intervening regions are variable in length and/or composition, thus providing the base composition “signature” (BCS) for each organism. Primer pairs were chosen so the total length of the amplified region is less than about 80-90 nucleotides. The label for each primer pair represents the starting and ending base number of the amplified region on the consensus diagram.

Included in the short bacterial database cross-section in Table 2 are many well known pathogens/biowarfare agents (shown in bold/red typeface) such as *Bacillus anthracis* or *Yersinia pestis* as well as some of the bacterial organisms found commonly in the natural environment such as *Streptomyces*. Even closely related organisms can be distinguished from each other by the appropriate choice of primers. For instance, two low G+C organisms, *Bacillus anthracis* and *Staph aureus*, can be distinguished from each other by using the primer pair defined by 16S_1337 or 23S_855 (ΔM of 4 Da).

Table 2: Cross Section Of A Database Of Calculated Molecular Masses¹

| Primer Regions ----> | 16S_971 | 16S_1100 | 16S_1337 | 16S_1294 | 16S_1228 | 23S_1021 | 23S_855 | 23S_193 | 23S_115 |
|------------------------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| Bug Name | | | | | | | | | |
| <i>Acinetobacter calcoaceticus</i> | 55619.1 | 55004 | 28446.7 | 35854.9 | 51295.4 | 30299 | 42654 | 39557.5 | 54999 |
| Bacillus anthracis | 55005 | 54388 | 28448 | 35238 | 51296 | 30295 | 42651 | 39560 | 56850 |
| <i>Bacillus cereus</i> | 55622.1 | 54387.9 | 28447.6 | 35854.9 | 51296.4 | 30295 | 42651 | 39560.5 | 56850.3 |
| <i>Bordetella bronchiseptica</i> | 56857.3 | 51300.4 | 28446.7 | 35857.9 | 51307.4 | 30299 | 42653 | 39559.5 | 51920.5 |
| <i>Borrelia burgdorferi</i> | 56231.2 | 55621.1 | 28440.7 | 35852.9 | 51295.4 | 30297 | 42029.9 | 38941.4 | 52524.6 |
| Brucella abortus | 58098 | 55011 | 28448 | 35854 | 50683 | | | | |
| <i>Campylobacter jejuni</i> | 58088.5 | 54386.9 | 29061.8 | 35856.9 | 50674.3 | 30294 | 42032.9 | 39558.5 | 45732.5 |
| <i>Chlamydia pneumoniae</i> | 55000 | 55007 | 29063 | 35855 | 50676 | 30295 | 42036 | 38941 | 56230 |
| Clostridium botulinum | 55006 | 53767 | 28445 | 35855 | 51291 | 30300 | 42656 | 39562 | 54999 |
| <i>Clostridium difficile</i> | 56855.3 | 54386.9 | 28444.7 | 35853.9 | 51296.4 | 30294 | 41417.8 | 39556.5 | 55612.2 |
| <i>Enterococcus faecalis</i> | 55620.1 | 54387.9 | 28447.6 | 35858.9 | 51296.4 | 30297 | 42652 | 39559.5 | 56849.3 |
| Escherichia coli | 55622 | 55009 | 28445 | 35857 | 51301 | 30301 | 42656 | 39562 | 54999 |
| <i>Francisella tularensis</i> | 53769 | 54385 | 28445 | 35856 | 51298 | | | | |
| <i>Haemophilus influenzae</i> | 55620.1 | 55006 | 28444.7 | 35855.9 | 51298.4 | 30298 | 42656 | 39560.5 | 55613.1 |
| <i>Klebsiella pneumoniae</i> | 55622.1 | 55008 | 28442.7 | 35856.9 | 51297.4 | 30300 | 42655 | 39562.5 | 55000 |
| Legionella pneumophila | 55618 | 55626 | 28446 | 35857 | 51303 | | | | |
| <i>Mycobacterium avium</i> | 54390.9 | 55631.1 | 29064.8 | 35858.9 | 51915.5 | 30298 | 42656 | 38942.4 | 56241.2 |
| <i>Mycobacterium leprae</i> | 54389.9 | 55629.1 | 29064.8 | 35860.9 | 51917.5 | 30298 | 42656 | 39559.5 | 56240.2 |
| <i>Mycobacterium tuberculosis</i> | 54390.9 | 55629.1 | 29064.8 | 35860.9 | 51301.4 | 30299 | 42656 | 39560.5 | 56243.2 |
| <i>Mycoplasma genitalium</i> | 53143.7 | 45115.4 | 29061.8 | 35854.9 | 50671.3 | 30294 | 43264.1 | 39558.5 | 56842.4 |
| <i>Mycoplasma pneumoniae</i> | 53143.7 | 45118.4 | 29061.8 | 35854.9 | 50673.3 | 30294 | 43264.1 | 39559.5 | 56843.4 |
| <i>Neisseria gonorrhoeae</i> | 55627.1 | 54389.9 | 28445.7 | 35855.9 | 51302.4 | 30300 | 42649 | 39561.5 | 55000 |
| Pseudomonas aeruginosa | 55623 | 55010 | 28443 | 35858 | 51301 | 30298 | 43272 | 39558 | 55619 |
| Rickettsia prowazekii | 58093 | 55621 | 28448 | 35853 | 50677 | 30293 | 42650 | 39559 | 53139 |
| Rickettsia rickettsii | 58094 | 55623 | 28448 | 35853 | 50679 | 30293 | 42648 | 39559 | 53755 |
| <i>Salmonella typhimurium</i> | 55622 | 55005 | 28445 | 35857 | 51301 | 30301 | 42658 | | |
| Shigella dysenteriae | 55623 | 55009 | 28444 | 35857 | 51301 | | | | |
| <i>Staphylococcus aureus</i> | 56854.3 | 54386.9 | 28443.7 | 35852.9 | 51294.4 | 30298 | 42655 | 39559.5 | 57466.4 |
| <i>Streptomyces</i> | 54389.9 | 59341.6 | 29063.8 | 35858.9 | 51300.4 | | | 39563.5 | 56864.3 |
| <i>Treponema pallidum</i> | 56245.2 | 55631.1 | 28445.7 | 35851.9 | 51297.4 | 30299 | 42034.9 | 38939.4 | 57473.4 |
| Vibrio cholerae | 55625 | 55626 | 28443 | 35857 | 52536 | 29063 | 30303 | 35241 | 50675 |
| <i>Vibrio parahaemolyticus</i> | 54384.9 | 55626.1 | 28444.7 | 34620.7 | 50064.2 | | | | |
| Yersinia pestis | 55620 | 55626 | 28443 | 35857 | 51299 | | | | |

¹Molecular mass distribution of PCR amplified regions for a selection of organisms (rows) across various primer pairs (columns). Pathogens are shown in **bold**. Empty cells indicate presently incomplete or missing data.

Figure 6 shows the use of ESI-FT-ICR MS for measurement of exact mass. The spectra from 46mer PCR products originating at position 1337 of the 16S rRNA from *S. aureus* (upper) and *B. anthracis* (lower) are shown. These data are from the region of the spectrum containing signals from the [M-8H]⁸⁺ charge states of the respective 5'-3' strands. The two strands differ by two (AT→CG) substitutions, and have measured masses of 14206.396 and 14208.373 ± 0.010 Da, respectively. The possible base compositions derived from the masses of the forward and reverse strands for the *B. anthracis* products are listed in Table 3.

Table 3: Possible base composition for *B. anthracis* products

| Calc. Mass | Error | Base Comp. |
|------------|----------|----------------|
| 14208.2935 | 0.079520 | A1 G17 C10 T18 |
| 14208.3160 | 0.056980 | A1 G20 C15 T10 |
| 14208.3386 | 0.034440 | A1 G23 C20 T2 |

| | | |
|-------------------|-----------------|------------------------|
| 14208.3074 | 0.065560 | A6 G11 C3 T26 |
| 14208.3300 | 0.043020 | A6 G14 C8 T18 |
| 14208.3525 | 0.020480 | A6 G17 C13 T10 |
| 14208.3751 | 0.002060 | A6 G20 C18 T2 |
| 14208.3439 | 0.029060 | A11 G8 C1 T26 |
| 14208.3665 | 0.006520 | A11 G11 C6 T18 |
| 14208.3890 | 0.016020 | A11 G14 C11 T10 |
| 14208.4116 | 0.038560 | A11 G17 C16 T2 |
| 14208.4030 | 0.029980 | A16 G8 C4 T18 |
| 14208.4255 | 0.052520 | A16 G11 C9 T10 |
| 14208.4481 | 0.075060 | A16 G14 C14 T2 |
| 14208.4395 | 0.066480 | A21 G5 C2 T18 |
| 14208.4620 | 0.089020 | A21 G8 C7 T10 |
| 14079.2624 | 0.080600 | A0 G14 C13 T19 |
| 14079.2849 | 0.058060 | A0 G17 C18 T11 |
| 14079.3075 | 0.035520 | A0 G20 C23 T3 |
| 14079.2538 | 0.089180 | A5 G5 C1 T35 |
| 14079.2764 | 0.066640 | A5 G8 C6 T27 |
| 14079.2989 | 0.044100 | A5 G11 C11 T19 |
| 14079.3214 | 0.021560 | A5 G14 C16 T11 |
| 14079.3440 | 0.000980 | A5 G17 C21 T3 |
| 14079.3129 | 0.030140 | A10 G5 C4 T27 |
| 14079.3354 | 0.007600 | A10 G8 C9 T19 |
| 14079.3579 | 0.014940 | A10 G11 C14 T11 |
| 14079.3805 | 0.037480 | A10 G14 C19 T3 |
| 14079.3494 | 0.006360 | A15 G2 C2 T27 |
| 14079.3719 | 0.028900 | A15 G5 C7 T19 |
| 14079.3944 | 0.051440 | A15 G8 C12 T11 |
| 14079.4170 | 0.073980 | A15 G11 C17 T3 |
| 14079.4084 | 0.065400 | A20 G2 C5 T19 |
| 14079.4309 | 0.087940 | A20 G5 C10 T13 |

Among the 16 compositions for the forward strand and the 18 compositions for the reverse strand that were calculated, only one pair (shown in **bold**) are complementary, corresponding to the actual base compositions of the *B. anthracis* PCR products.

5 **Example 4: BCS of Region from *Bacillus anthracis* and *Bacillus cereus***

A conserved *Bacillus* region from *B. anthracis* (A₁₄G₉C₁₄T₉) and *B. cereus* (A₁₅G₉C₁₃T₉) having a C to A base change was synthesized and subjected to ESI-TOF MS. The results are shown in Figure 7 in which the two regions are clearly distinguished using the method of the present invention (MW=14072.26 vs.
10 14096.29).

Example 5: Identification of additional bioagents

In other examples of the present invention, the pathogen *Vibrio cholera* can be distinguished from *Vibrio parahemolyticus* with $\Delta M > 600$ Da using one of three
15 16S primer sets shown in Table 2 (16S_971, 16S_1228 or 16S_1294) as shown in Table 4. The two mycoplasma species in the list (*M. genitalium* and *M. pneumoniae*) can also be distinguished from each other, as can the three mycobacteriae. While the direct mass measurements of amplified products can identify and distinguish a large number of organisms, measurement of the base composition signature provides
20 dramatically enhanced resolving power for closely related organisms. In cases such as *Bacillus anthracis* and *Bacillus cereus* that are virtually indistinguishable from each other based solely on mass differences, compositional analysis or fragmentation patterns are used to resolve the differences. The single base difference between the two organisms yields different fragmentation patterns, and despite the presence of the
25 ambiguous/unidentified base N at position 20 in *B. anthracis*, the two organisms can be identified.

Tables 4a-b show examples of primer pairs from Table 1 which distinguish pathogens from background.

Table 4a

| Organism name | 23S_855 | 16S_1337 | 23S_1021 |
|------------------------------|----------|----------|----------|
| <i>Bacillus anthracis</i> | 42650.98 | 28447.65 | 30294.98 |
| <i>Staphylococcus aureus</i> | 42654.97 | 28443.67 | 30297.96 |

Table 4b

| Organism name | 16S_971 | 16S_1294 | 16S_1228 |
|--------------------------------|----------|----------|----------|
| <i>Vibrio cholerae</i> | 55625.09 | 35856.87 | 52535.59 |
| <i>Vibrio parahaemolyticus</i> | 54384.91 | 34620.67 | 50064.19 |

Table 4 shows the expected molecular weight and base composition of region 16S_1100-1188 in *Mycobacterium avium* and *Streptomyces sp.*

5

Table 5

| Region | Organism name | Length | Molecular weight | Base comp. |
|---------------|----------------------------|--------|------------------|---|
| 16S_1100-1188 | <i>Mycobacterium avium</i> | 82 | 25624.1728 | A ₁₆ G ₃₂ C ₁₈ T ₁₆ |
| 16S_1100-1188 | <i>Streptomyces sp.</i> | 96 | 29904.871 | A ₁₇ G ₃₈ C ₂₇ T ₁₄ |

Table 5 shows base composition (single strand) results for 16S_1100-1188 primer amplification reactions different species of bacteria. Species which are repeated in the table (e.g., *Clostridium botulinum*) are different strains which have different base compositions in the 16S_1100-1188 region.

10

Table 6

| Organism name | Base comp. | Organism name | Base comp. |
|---------------------------------|---|----------------------------------|---|
| <i>Mycobacterium avium</i> | A ₁₆ G ₃₂ C ₁₈ T ₁₆ | <i>Vibrio cholerae</i> | A ₂₃ G ₃₀ C ₂₁ T ₁₆ |
| <i>Streptomyces sp.</i> | A ₁₇ G ₃₈ C ₂₇ T ₁₄ | <i>Aeromonas hydrophila</i> | A ₂₃ G ₃₁ C ₂₁ T ₁₅ |
| <i>Ureaplasma urealyticum</i> | A ₁₈ G ₃₀ C ₁₇ T ₁₇ | <i>Aeromonas salmonicida</i> | A ₂₃ G ₃₁ C ₂₁ T ₁₅ |
| <i>Streptomyces sp.</i> | A ₁₉ G ₃₆ C ₂₄ T ₁₈ | <i>Mycoplasma genitalium</i> | A ₂₄ G ₁₉ C ₁₂ T ₁₈ |
| <i>Mycobacterium leprae</i> | A ₂₀ G ₃₂ C ₂₂ T ₁₆ | <i>Clostridium botulinum</i> | A ₂₄ G ₂₅ C ₁₈ T ₂₀ |
| <i>M. tuberculosis</i> | A ₂₀ G ₃₃ C ₂₁ T ₁₆ | <i>Bordetella bronchiseptica</i> | A ₂₄ G ₂₆ C ₁₉ T ₁₄ |
| <i>Nocardia asteroides</i> | A ₂₀ G ₃₃ C ₂₁ T ₁₆ | <i>Francisella tularensis</i> | A ₂₄ G ₂₆ C ₁₉ T ₁₉ |
| <i>Fusobacterium necroforum</i> | A ₂₁ G ₂₆ C ₂₂ T ₁₈ | <i>Bacillus anthracis</i> | A ₂₄ G ₂₆ C ₂₀ T ₁₈ |
| <i>Listeria monocytogenes</i> | A ₂₁ G ₂₇ C ₁₉ T ₁₉ | <i>Campylobacter jejuni</i> | A ₂₄ G ₂₆ C ₂₀ T ₁₈ |
| <i>Clostridium botulinum</i> | A ₂₁ G ₂₇ C ₁₉ T ₂₁ | <i>Staphylococcus aureus</i> | A ₂₄ G ₂₆ C ₂₀ T ₁₈ |
| <i>Neisseria gonorrhoeae</i> | A ₂₁ G ₂₈ C ₂₁ T ₁₈ | <i>Helicobacter pylori</i> | A ₂₄ G ₂₆ C ₂₀ T ₁₉ |
| <i>Bartonella quintana</i> | A ₂₁ G ₃₀ C ₂₂ T ₁₆ | <i>Helicobacter pylori</i> | A ₂₄ G ₂₆ C ₂₁ T ₁₈ |
| <i>Enterococcus faecalis</i> | A ₂₂ G ₂₇ C ₂₀ T ₁₉ | <i>Moraxella catarrhalis</i> | A ₂₄ G ₂₆ C ₂₃ T ₁₆ |

| | | | |
|---------------------------------|---|--|---|
| <i>Bacillus megaterium</i> | A ₂₂ G ₂₈ C ₂₀ T ₁₈ | <i>Haemophilus influenzae</i> <i>Rd</i> | A ₂₄ G ₂₈ C ₂₀ T ₁₇ |
| <i>Bacillus subtilis</i> | A ₂₂ G ₂₈ C ₂₁ T ₁₇ | <i>Chlamydia trachomatis</i> | A ₂₄ G ₂₈ C ₂₁ T ₁₆ |
| <i>Pseudomonas aeruginosa</i> | A ₂₂ G ₂₉ C ₂₃ T ₁₅ | <i>Chlamydophila pneumoniae</i> | A ₂₄ G ₂₈ C ₂₁ T ₁₆ |
| <i>Legionella pneumophila</i> | A ₂₂ G ₃₂ C ₂₀ T ₁₆ | <i>C. pneumonia AR39</i> | A ₂₄ G ₂₈ C ₂₁ T ₁₆ |
| <i>Mycoplasma pneumoniae</i> | A ₂₃ G ₂₀ C ₁₄ T ₁₆ | <i>Pseudomonas putida</i> | A ₂₄ G ₂₉ C ₂₁ T ₁₆ |
| <i>Clostridium botulinum</i> | A ₂₃ G ₂₆ C ₂₀ T ₁₉ | <i>Proteus vulgaris</i> | A ₂₄ G ₃₀ C ₂₁ T ₁₅ |
| <i>Enterococcus faecium</i> | A ₂₃ G ₂₆ C ₂₁ T ₁₈ | <i>Yersinia pestis</i> | A ₂₄ G ₃₀ C ₂₁ T ₁₅ |
| <i>Acinetobacter calcoaceti</i> | A ₂₃ G ₂₆ C ₂₁ T ₁₉ | <i>Yersinia pseudotuberculosis</i> | A ₂₄ G ₃₀ C ₂₁ T ₁₅ |
| <i>Leptospira borgpeterseni</i> | A ₂₃ G ₂₆ C ₂₄ T ₁₅ | <i>Clostridium botulinum</i> | A ₂₅ G ₂₄ C ₁₈ T ₂₁ |
| <i>Leptospira interrogans</i> | A ₂₃ G ₂₆ C ₂₄ T ₁₅ | <i>Clostridium tetani</i> | A ₂₅ G ₂₅ C ₁₈ T ₂₀ |
| <i>Clostridium perfringens</i> | A ₂₃ G ₂₇ C ₁₉ T ₁₉ | <i>Francisella tularensis</i> | A ₂₅ G ₂₅ C ₁₉ T ₁₉ |
| <i>Bacillus anthracis</i> | A ₂₃ G ₂₇ C ₂₀ T ₁₈ | <i>Acinetobacter calcoacetic</i> | A ₂₅ G ₂₆ C ₂₀ T ₁₉ |
| <i>Bacillus cereus</i> | A ₂₃ G ₂₇ C ₂₀ T ₁₈ | <i>Bacteriodes fragilis</i> | A ₂₅ G ₂₇ C ₁₆ T ₂₂ |
| <i>Bacillus thuringiensis</i> | A ₂₃ G ₂₇ C ₂₀ T ₁₈ | <i>Chlamydophila psittaci</i> | A ₂₅ G ₂₇ C ₂₁ T ₁₆ |
| <i>Aeromonas hydrophila</i> | A ₂₃ G ₂₉ C ₂₁ T ₁₆ | <i>Borrelia burgdorferi</i> | A ₂₅ G ₂₉ C ₁₇ T ₁₉ |
| <i>Escherichia coli</i> | A ₂₃ G ₂₉ C ₂₁ T ₁₆ | <i>Streptobacillus monilifor</i> | A ₂₆ G ₂₆ C ₂₀ T ₁₆ |
| <i>Pseudomonas putida</i> | A ₂₃ G ₂₉ C ₂₁ T ₁₇ | <i>Rickettsia prowazekii</i> | A ₂₆ G ₂₈ C ₁₈ T ₁₈ |
| <i>Escherichia coli</i> | A ₂₃ G ₂₉ C ₂₂ T ₁₅ | <i>Rickettsia rickettsii</i> | A ₂₆ G ₂₈ C ₂₀ T ₁₆ |
| <i>Shigella dysenteriae</i> | A ₂₃ G ₂₉ C ₂₂ T ₁₅ | <i>Mycoplasma mycoides</i> | A ₂₈ G ₂₃ C ₁₆ T ₂₀ |

The same organism having different base compositions are different strains. Groups of organisms which are highlighted or in italics have the same base compositions in the amplified region. Some of these organisms can be distinguished using multiple primers. For example, *Bacillus anthracis* can be distinguished from *Bacillus cereus* and *Bacillus thuringiensis* using the primer 16S_971-1062 (Table 6). Other primer pairs which produce unique base composition signatures are shown in Table 6 (bold). Clusters containing very similar threat and ubiquitous non-threat organisms (e.g. *anthracis* cluster) are distinguished at high resolution with focused sets of primer pairs. The known biowarfare agents in Table 6 are *Bacillus anthracis*, *Yersinia pestis*, *Francisella tularensis* and *Rickettsia prowazekii*.

Table 7

| Organism | 16S_971-1062 | 16S_1228-1310 | 16S_1100-1188 |
|------------------------------------|---|---|---|
| <i>Aeromonas hydrophila</i> | A ₂₁ G ₂₉ C ₂₂ T ₂₀ | A ₂₂ G ₂₇ C ₂₁ T ₁₃ | A ₂₃ G ₃₁ C ₂₁ T ₁₅ |
| <i>Aeromonas salmonicida</i> | A ₂₁ G ₂₉ C ₂₂ T ₂₀ | A ₂₂ G ₂₇ C ₂₁ T ₁₃ | A ₂₃ G ₃₁ C ₂₁ T ₁₅ |
| <i>Bacillus anthracis</i> | A₂₁G₂₇C₂₂T₂₂ | A ₂₄ G ₂₂ C ₁₉ T ₁₈ | A ₂₃ G ₂₇ C ₂₀ T ₁₈ |
| <i>Bacillus cereus</i> | A ₂₂ G ₂₇ C ₂₁ T ₂₂ | A ₂₄ G ₂₂ C ₁₉ T ₁₈ | A ₂₃ G ₂₇ C ₂₀ T ₁₈ |
| <i>Bacillus thuringiensis</i> | A ₂₂ G ₂₇ C ₂₁ T ₂₂ | A ₂₄ G ₂₂ C ₁₉ T ₁₈ | A ₂₃ G ₂₇ C ₂₀ T ₁₈ |
| <i>Chlamydia trachomatis</i> | A₂₂G₂₆C₂₀T₂₃ | A₂₄G₂₃C₁₉T₁₆ | A ₂₄ G ₂₈ C ₂₁ T ₁₆ |
| <i>Chlamydia pneumoniae</i> AR39 | A ₂₆ G ₂₃ C ₂₀ T ₂₂ | A ₂₆ G ₂₂ C ₁₆ T ₁₈ | A ₂₄ G ₂₈ C ₂₁ T ₁₆ |
| <i>Leptospira borgpetersenii</i> | A ₂₂ G ₂₆ C ₂₀ T ₂₁ | A ₂₂ G ₂₅ C ₂₁ T ₁₅ | A ₂₃ G ₂₆ C ₂₄ T ₁₅ |
| <i>Leptospira interrogans</i> | A ₂₂ G ₂₆ C ₂₀ T ₂₁ | A ₂₂ G ₂₅ C ₂₁ T ₁₅ | A ₂₃ G ₂₆ C ₂₄ T ₁₅ |
| <i>Mycoplasma genitalium</i> | A ₂₈ G ₂₃ C ₁₅ T ₂₂ | A₃₀G₁₈C₁₅T₁₉ | A₂₄G₁₉C₁₂T₁₈ |
| <i>Mycoplasma pneumoniae</i> | A ₂₈ G ₂₃ C ₁₅ T ₂₂ | A₂₇G₁₉C₁₆T₂₀ | A₂₃G₂₀C₁₄T₁₆ |
| <i>Escherichia coli</i> | A₂₂G₂₈C₂₀T₂₂ | A ₂₄ G ₂₅ C ₂₁ T ₁₃ | A ₂₃ G ₂₉ C ₂₂ T ₁₅ |
| <i>Shigella dysenteriae</i> | A₂₂G₂₈C₂₁T₂₁ | A ₂₄ G ₂₅ C ₂₁ T ₁₃ | A ₂₃ G ₂₉ C ₂₂ T ₁₅ |
| <i>Proteus vulgaris</i> | A₂₃G₂₆C₂₂T₂₁ | A₂₆G₂₄C₁₉T₁₄ | A ₂₄ G ₃₀ C ₂₁ T ₁₅ |
| <i>Yersinia pestis</i> | A ₂₄ G ₂₅ C ₂₁ T ₂₂ | A ₂₅ G ₂₄ C ₂₀ T ₁₄ | A ₂₄ G ₃₀ C ₂₁ T ₁₅ |
| <i>Yersinia pseudotuberculosis</i> | A ₂₄ G ₂₅ C ₂₁ T ₂₂ | A ₂₅ G ₂₄ C ₂₀ T ₁₄ | A ₂₄ G ₃₀ C ₂₁ T ₁₅ |
| <i>Francisella tularensis</i> | A₂₀G₂₅C₂₁T₂₃ | A₂₃G₂₆C₁₇T₁₇ | A₂₄G₂₆C₁₉T₁₉ |
| <i>Rickettsia prowazekii</i> | A₂₁G₂₆C₂₄T₂₅ | A₂₄G₂₃C₁₆T₁₉ | A₂₆G₂₈C₁₈T₁₈ |
| <i>Rickettsia rickettsii</i> | A₂₁G₂₆C₂₅T₂₄ | A₂₄G₂₄C₁₇T₁₇ | A₂₆G₂₈C₂₀T₁₆ |

The sequence of *B. anthracis* and *B. cereus* in region 16S_971 is shown below. Shown in bold is the single base difference between the two species which can be detected using the methods of the present invention. *B. anthracis* has an ambiguous base at position 20.

*B. anthracis*_16S_971

GCGAAGAACCUUACCAGGUMUUGACAUCCUCUGACAACCCUAGAGAUAG
GGCUUCUCCUUCGGGAGCAGAGUGACAGGUGGUGCAUGGUU (SEQ ID
NO:4)

*B.cereus*_16S_971

GCGAAGAACCUUACCAGGUCUUGACAUCCUCUGAAAACCCUAGAGAUAG
GGCUUCUCCUUCGGGAGCAGAGUGACAGGUGGUGCAUGGUU (SEQ ID
NO:5)

5

Example 6: ESI-TOF MS of sspE 56-mer Plus Calibrant

The mass measurement accuracy that can be obtained using an internal mass standard in the ESI-MS study of PCR products is shown in Fig.8. The mass standard was a 20-mer phosphorothioate oligonucleotide added to a solution containing a 56-mer PCR product from the *B. anthracis* spore coat protein sspE. The mass of the expected PCR product distinguishes *B. anthracis* from other species of *Bacillus* such as *B. thuringiensis* and *B. cereus*.

Example 7: *B. anthracis* ESI-TOF Synthetic 16S_1228 Duplex

An ESI-TOF MS spectrum was obtained from an aqueous solution containing 5 μ M each of synthetic analogs of the expected forward and reverse PCR products from the nucleotide 1228 region of the *B. anthracis* 16S rRNA gene. The results (Fig. 9) show that the molecular weights of the forward and reverse strands can be accurately determined and easily distinguish the two strands. The $[M-21H^+]^{21-}$ and $[M-20H^+]^{20-}$ charge states are shown.

Example 8: ESI-FTICR-MS of Synthetic *B. anthracis* 16S_1337 46 Base Pair Duplex

An ESI-FTICR-MS spectrum was obtained from an aqueous solution containing 5 μ M each of synthetic analogs of the expected forward and reverse PCR products from the nucleotide 1337 region of the *B. anthracis* 16S rRNA gene. The results (Fig. 10) show that the molecular weights of the strands can be distinguished by this method. The $[M-16H^+]^{16-}$ through $[M-10H^+]^{10-}$ charge states are shown. The insert highlights the resolution that can be realized on the FTICR-MS instrument, which allows the charge state of the ion to be determined from the mass difference between peaks differing by a single ^{13}C substitution.

Example 9: ESI-TOF MS of 56-mer Oligonucleotide from saspB Gene of *B. anthracis* with Internal Mass Standard

ESI-TOF MS spectra were obtained on a synthetic 56-mer oligonucleotide (5 μ M) from the saspB gene of *B. anthracis* containing an internal mass standard at an ESI of 1.7 μ L/min as a function of sample consumption. The results (Fig. 11) show that the signal to noise is improved as more scans are summed, and that the standard and the product are visible after only 100 scans.

Example 10: ESI-TOF MS of an Internal Standard with Tributylammonium (TBA)-trifluoroacetate (TFA) Buffer

An ESI-TOF-MS spectrum of a 20-mer phosphorothioate mass standard was obtained following addition of 5 mM TBA-TFA buffer to the solution. This buffer strips charge from the oligonucleotide and shifts the most abundant charge state from $[M-8H]^+{}^8$ to $[M-3H]^+{}^3$ (Fig. 12).

Example 11: Master Database Comparison

The molecular masses obtained through Examples 1-10 are compared to molecular masses of known bioagents stored in a master database to obtain a high probability matching molecular mass.

Example 12: Master Data Base Interrogation over the Internet

The same procedure as in Example 11 is followed except that the local computer did not store the Master database. The Master database is interrogated over an internet connection, searching for a molecular mass match.

Example 13: Master Database Updating

The same procedure as in example 11 is followed except the local computer is connected to the internet and has the ability to store a master database locally. The local computer system periodically, or at the user's discretion, interrogates the Master database, synchronizing the local master database with the global Master database. This provides the current molecular mass information to both the local database as well as to the global Master database. This further provides more of a globalized knowledge base.

Example 14: Global Database Updating

The same procedure as in example 13 is followed except there are numerous such local stations throughout the world. The synchronization of each database adds to the diversity of information and diversity of the molecular masses of known
5 bioagents.

Various modifications of the invention, in addition to those described herein, will be apparent to those skilled in the art from the foregoing description. Such modifications are also intended to fall within the scope of the appended claims.

10

What is claimed is:

1. A method of identifying an unknown bioagent using a database of molecular masses of known bioagents comprising:
 - contacting nucleic acid from said bioagent with at least one pair of
 - 5 oligonucleotide primers that hybridize to sequences of said nucleic acid, wherein said sequences flank a variable nucleic acid sequence of said bioagent;
 - producing an amplification product of said variable nucleic acid sequence;
 - determining a first molecular mass of said amplification product; and
 - comparing said first molecular mass to the molecular masses of known
 - 10 bioagents, thereby identifying the unknown bioagent.
2. The method of claim 1 wherein said sequences to which said at least one pair of oligonucleotide primers hybridize are highly conserved.
3. The method of claim 1 wherein said sequences to which said at least one pair of oligonucleotide primers hybridize are highly conserved across at least two species.
- 15 4. The method of claim 1 further comprising the step of isolating a nucleic acid from said bioagent prior to contacting said nucleic acid with said at least one pair of oligonucleotide primers, wherein the comparing step further comprises comparing a base-pair count resulting from a translation of the corresponding molecular mass, and wherein a master database of molecular masses of known bioagents further includes a
- 20 translation of said molecular masses of known bioagents to corresponding base-pair counts of each known bioagent resulting from a specific primer pair set and comparing the base-pair count of said unknown bioagent against the obtained base-pair count of known bioagents for the selected primer pair set for determining the identity of said unknown bioagent.
- 25 5. The method of claim 4 further comprising the step of reconciling the database of molecular masses of known bioagents with the master database of molecular masses of known bioagents.
6. The method of claim 1 wherein said bioagent is a bacterium, virus, cell or spore.
- 30 7. The method of claim 1 wherein said amplification product is ionized by electrospray ionization, matrix assisted laser desorption or fast atom bombardment prior to molecular mass determination.

8. The method of claim 1 wherein said molecular mass is determined by mass spectrometry.
9. The method of claim 5 wherein said master database of molecular masses of known bioagents and the database of molecular masses of known bioagents are reconciled over a network.
10. The method of claim 4 wherein the identity is determined by statistically correlating the molecular mass of the unknown bioagent with at least one molecular mass of said master database.
11. A database having cell-data positional significance comprising at least a first table of a plurality of data-containing cells, said first table organized into at least a first row and a second row, each row having columns and data-containing cells; and wherein said data-containing cells have an alignment with at least one other row for differentiating aligned from non-aligned data-containing cells, and wherein said differentiation in alignment of said data-containing cells designates a structural feature of a polymer.
12. The database according to claim 11 wherein said alignment is a vertical alignment according to base pair homology along a linear segment within each polymer.
13. The database according to claim 11 wherein said vertical alignment further aligns cell-data according to inter-species conserved regions.
14. The database according to claim 11 wherein the structural feature is a bulge or a loop.
15. The database according to claim 11 wherein the polymer is an RNA.
16. A method for reconciling a first file and a second file, said second file corresponding at least in part to said first file, said first file and said second file each containing records, said records corresponding to rows in a table of a dimensional database having rows and columns defined by data-cells having data-cell positional significance, said method comprising:
- comparing said first file and said corresponding second file with a backup file containing records from a previous reconciliation of said first file and said corresponding second file to identify new, updated or deleted records;
- creating a reconcile file containing information pertaining to said new, updated or deleted records identified in said comparing step; and

copying the contents of said reconcile file to said first file, said corresponding second file and a new backup file.

17. A service providing information related to a bioagent comprising:

providing a dimensional master database for storing a molecular mass, an identity and a detail corresponding to a plurality of known bioagents and, said master database storing the molecular mass, the identity and the detail for a plurality of known bioagents;

interrogating the master database with an identification request of an unknown bioagent to generate a response; and

10 delivering said response from the master database to a requester.

18. The service according to claim 17 wherein the molecular mass is of a selected portion of the known bioagent, the identity comprises at least a geographic origin and a name for the known bioagent, and the detail comprises at least a treatment.

19. The service according to claim 17 wherein the request comprises a symptomatology and the identification comprises a recommended pair of primers for hybridizing to sequences of nucleic acid flanking a variable nucleic acid sequence of the unknown bioagent, and said pair of primers are hybridized to the sequences of nucleic acid flanking a variable nucleic acid sequence of the unknown bioagent.

20. The service according to claim 19 wherein the nucleic acid sequence of the unknown bioagent between said pair of primers defines the selected portion of both the known bioagents and the unknown bioagent.

21. The service according to claim 20 wherein the response is delivered through a network.

22. The service according to claim 20 wherein the request comprises a molecular mass of the unknown bioagent for the selected portion and where the response generated thereto comprises a set of molecular masses for analogous selected portions of know bioagents, and said set comprising at least one molecular mass from the master database.

23. The service according to claim 21 wherein the network is a local area network.

24. The service according to claim 21 wherein the network is a wide area network.

25. The service according to claim 22 wherein the network is the internet.

26. A method of determining a geographical origin of a selected bioagent using a database of molecular masses of known bioagents comprising:
- contacting nucleic acid from said selected bioagent with at least one pair of oligonucleotide primers which hybridize to sequences of said nucleic acid, wherein
- 5 said sequences flank a variable nucleic acid sequence of the bioagent;
- producing an amplification product of said variable nucleic acid sequence;
- determining a first molecular mass of said amplification product; and
- comparing said first molecular mass to the molecular masses of known
- 10 bioagents for determining a geographic origin of said selected bioagent, said comparison determining an identity and a geographic origin of said selected bioagent.
27. The method of claim 26 wherein said sequences to which said at least one pair of oligonucleotide primers hybridize are highly conserved.
28. The method of claim 26 wherein said sequences to which said at least one pair of oligonucleotide primers hybridize are highly conserved across species.
- 15 29. The method of claim 26 further comprising the step of isolating a nucleic acid from said selected bioagent prior to contacting said nucleic acid with said at least one pair of oligonucleotide primers, wherein the comparing step further comprises interrogating a master database of molecular masses of known bioagents for obtaining molecular masses of known bioagents and comparing the molecular mass of said
- 20 selected bioagent against the obtained molecular masses of known bioagents thereby determining an origin of said selected bioagent.
30. The method of claim 29 further comprising the step of reconciling the database of molecular masses of known bioagents with the master database of molecular masses of known bioagents.
- 25 31. The method of claim 26 wherein said bioagent is a bacterium, virus, cell or spore.
32. The method of claim 26 wherein said amplification product is ionized by electrospray ionization, matrix assisted laser desorption or fast atom bombardment prior to molecular mass determination.
- 30 33. The method of claim 26 wherein said molecular mass is determined by mass spectrometry.
34. The method of claim 29 wherein said master database of molecular masses of known bioagents and the database of molecular masses of known bioagents are reconciled over a network.

35. The method of claim 29 wherein the origin comprises a statistical group of matching molecular masses and the geographic origin corresponding thereto.

Fig. 1A

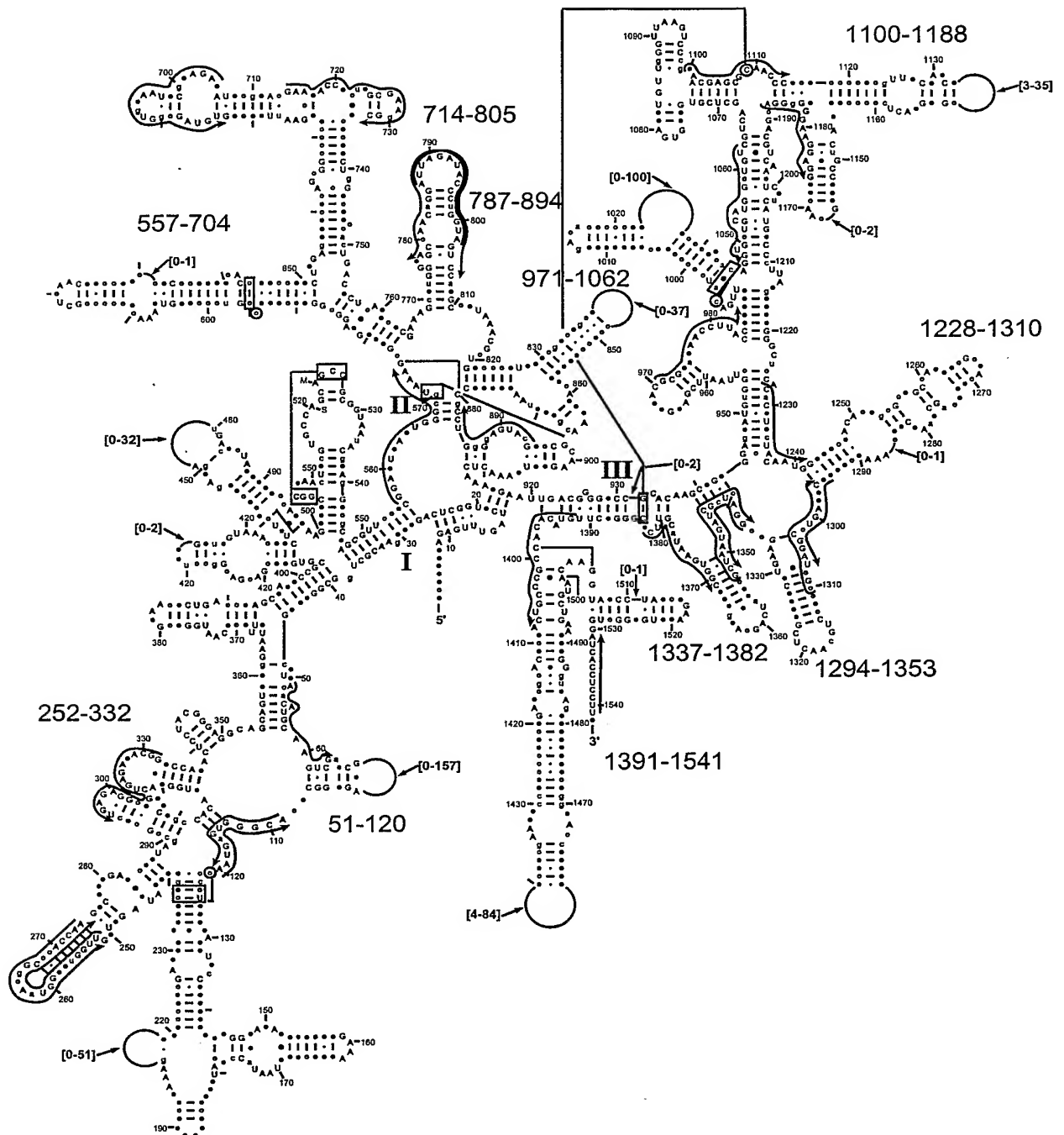


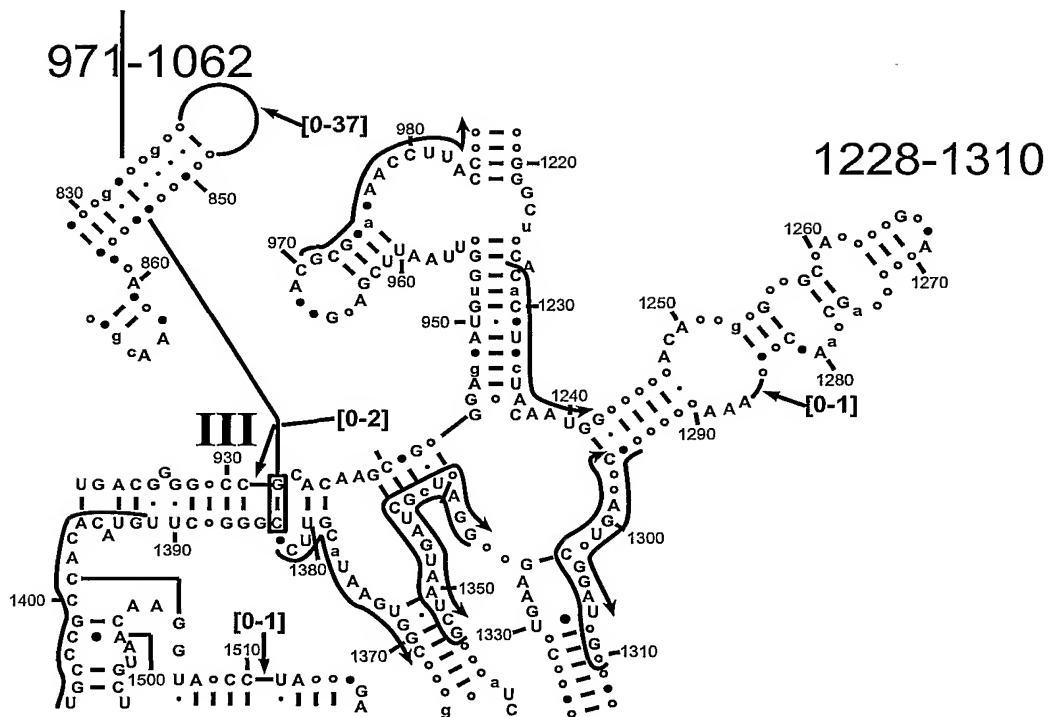
Fig. 1A-2

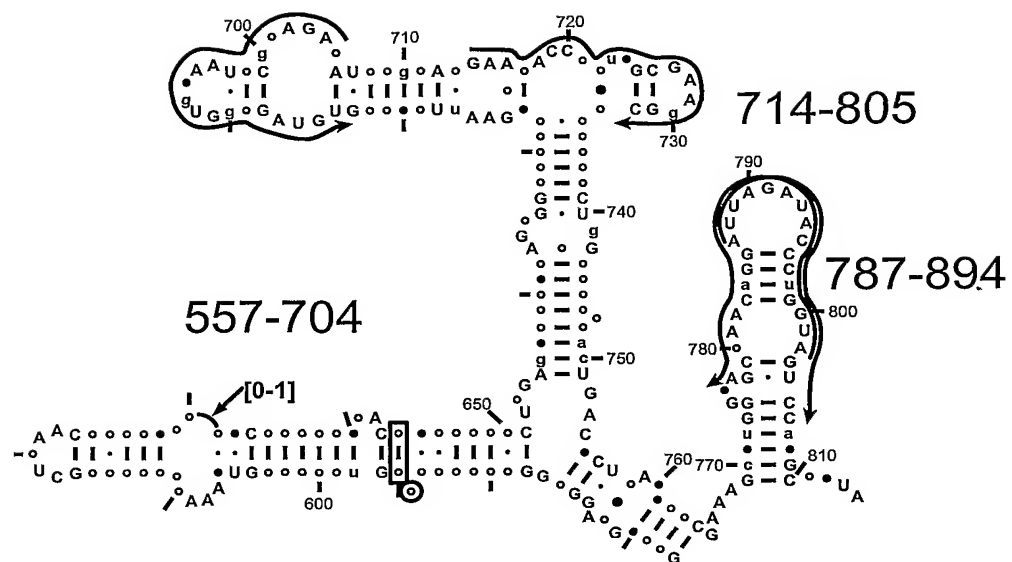
Fig. 1A-3

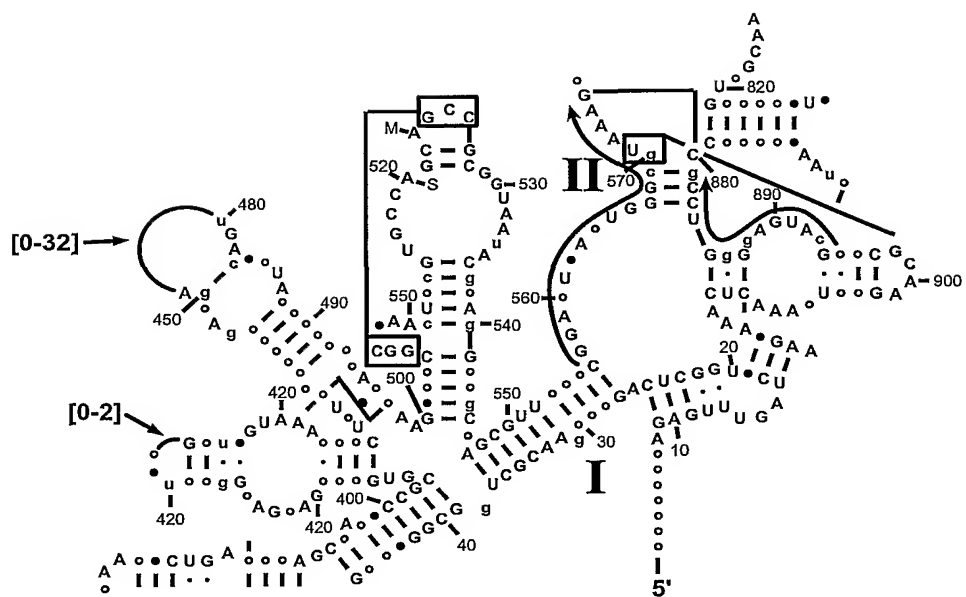
Fig. 1A-4

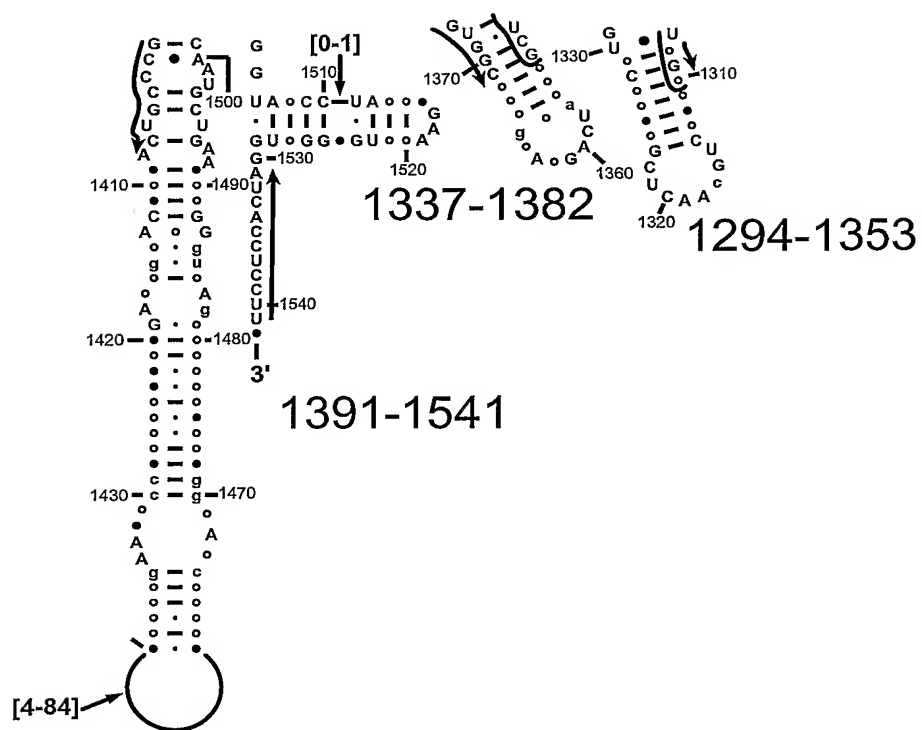
Fig. 1B-1

Fig. 1B-2

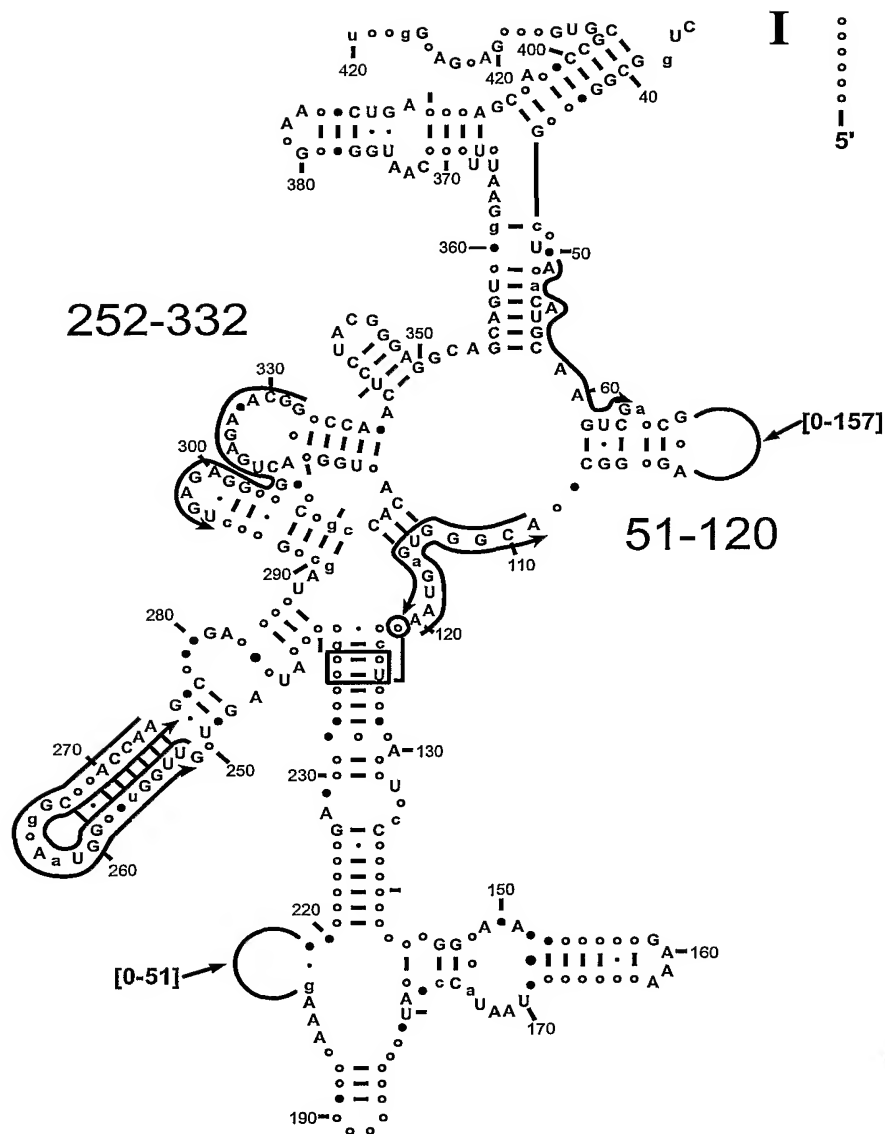


Fig. 1C-1

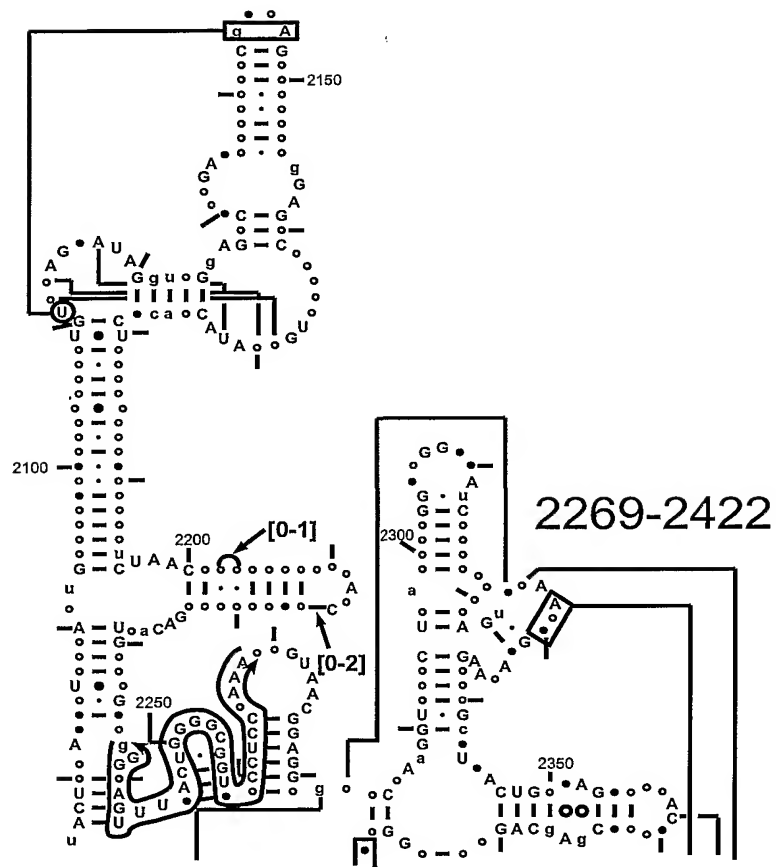


Fig. 1C-2

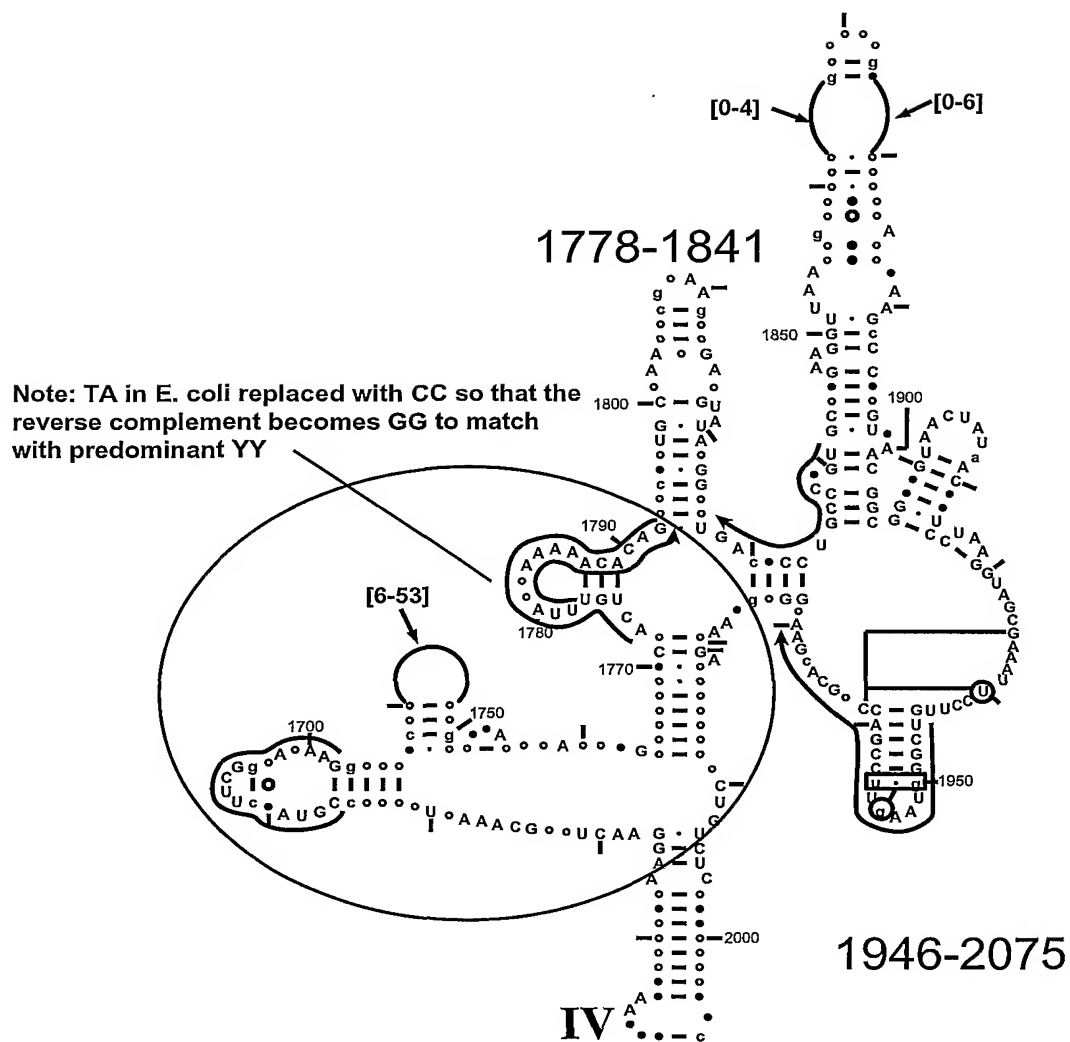


Fig. 1D

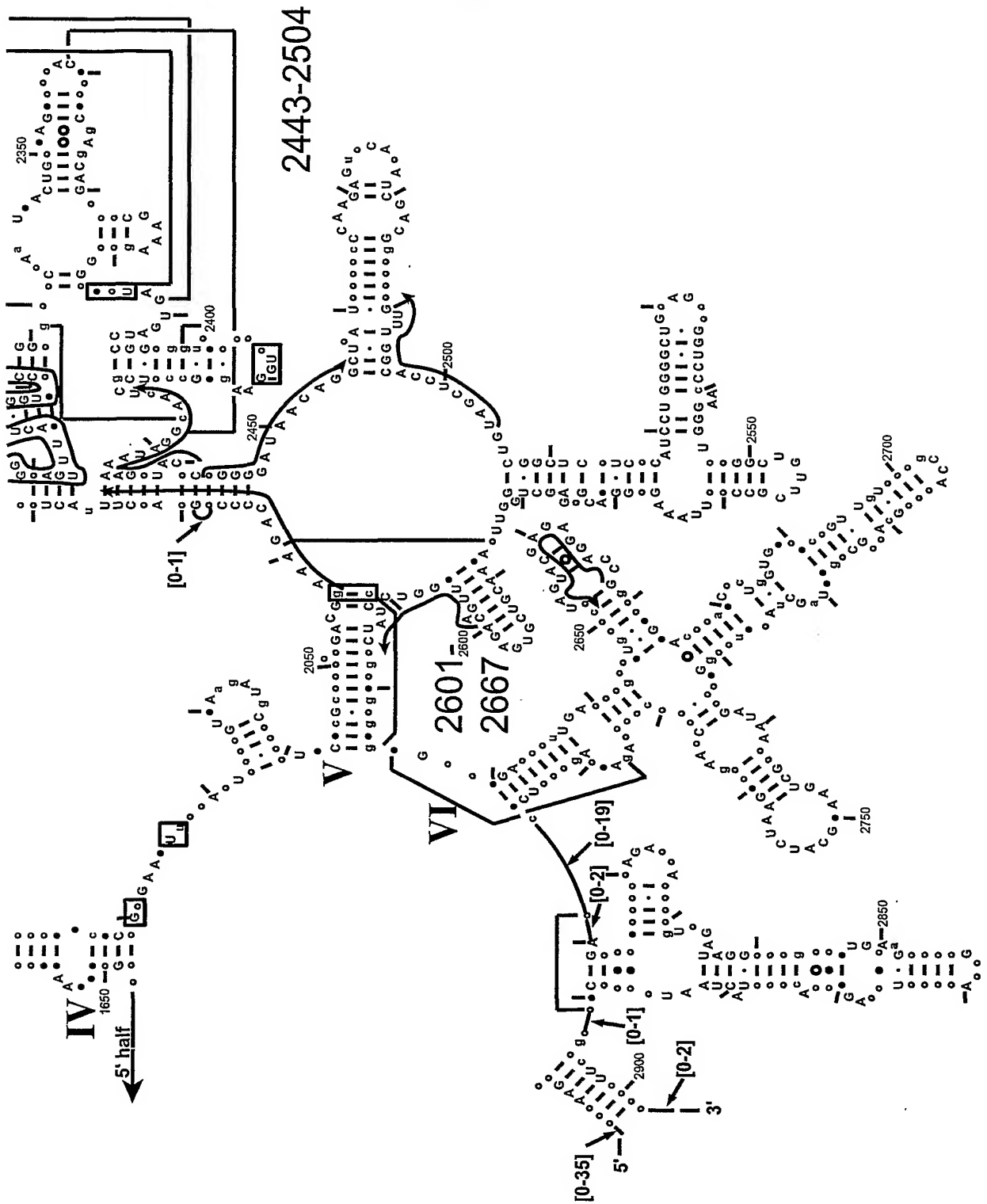


Fig. 1E

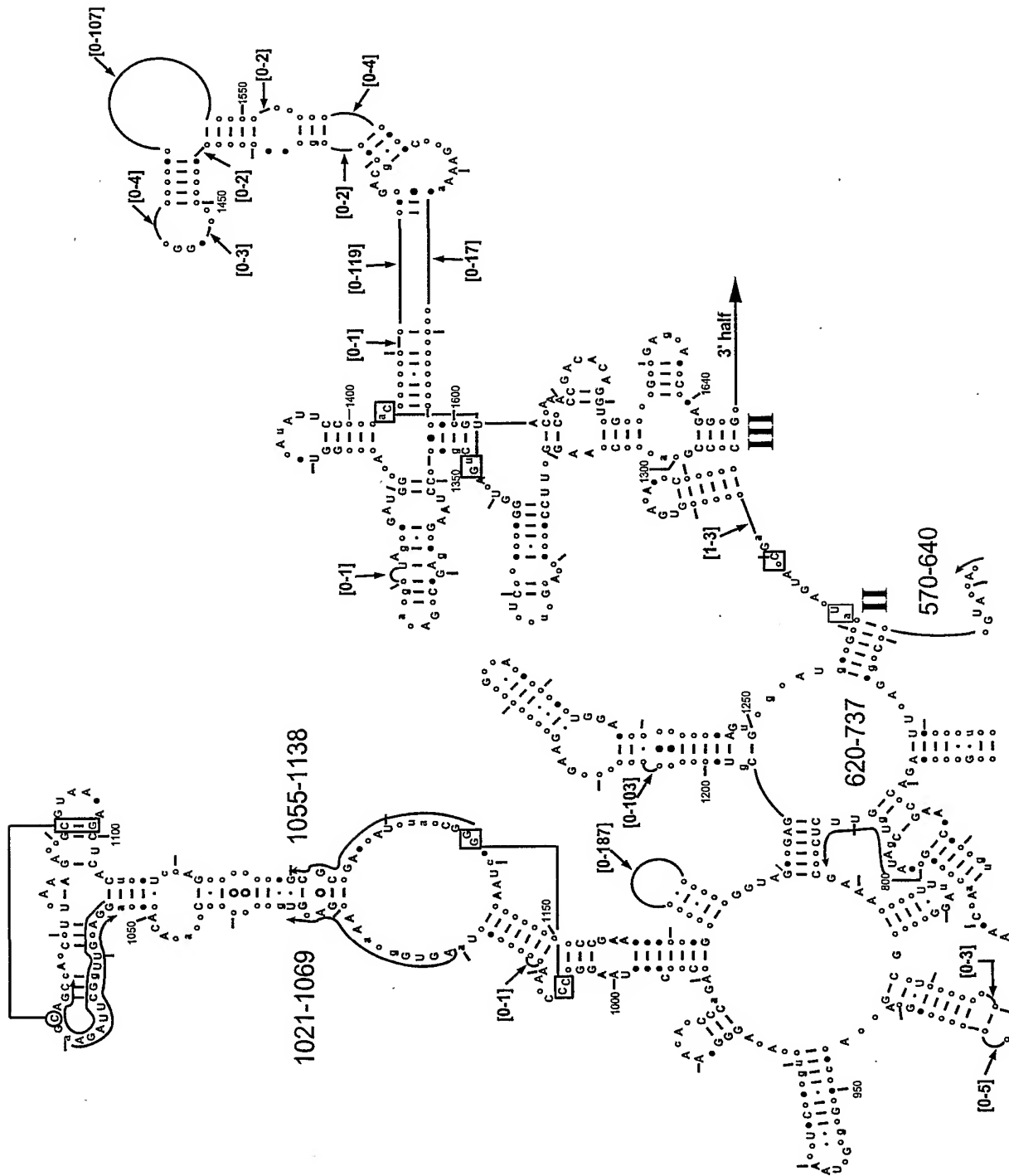


Fig. 1F

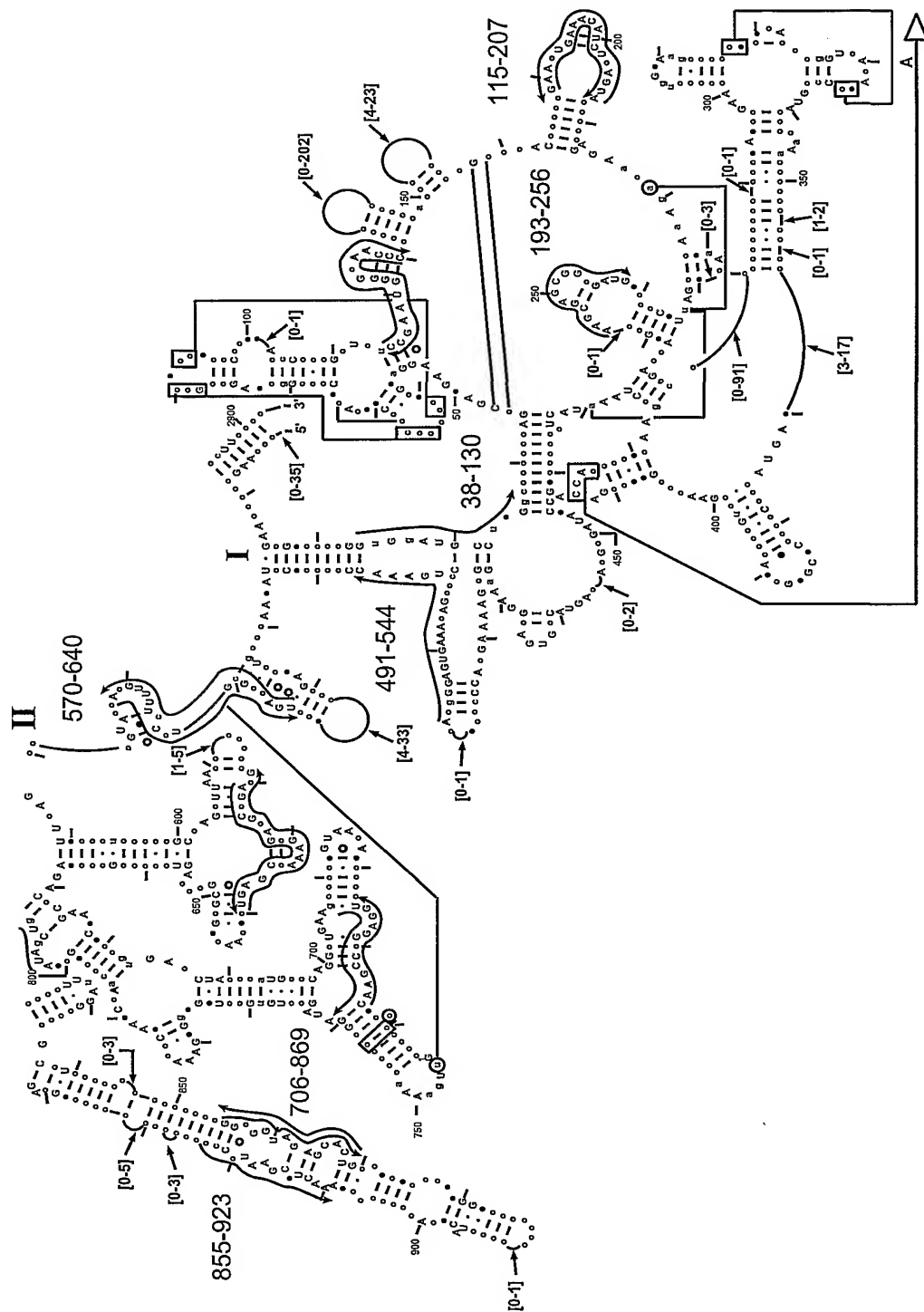


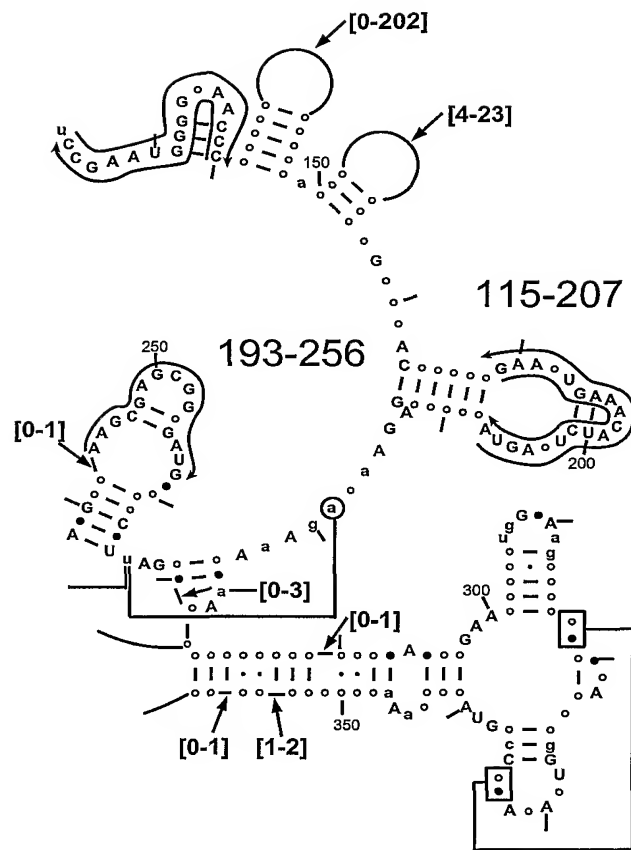
Fig. 1G

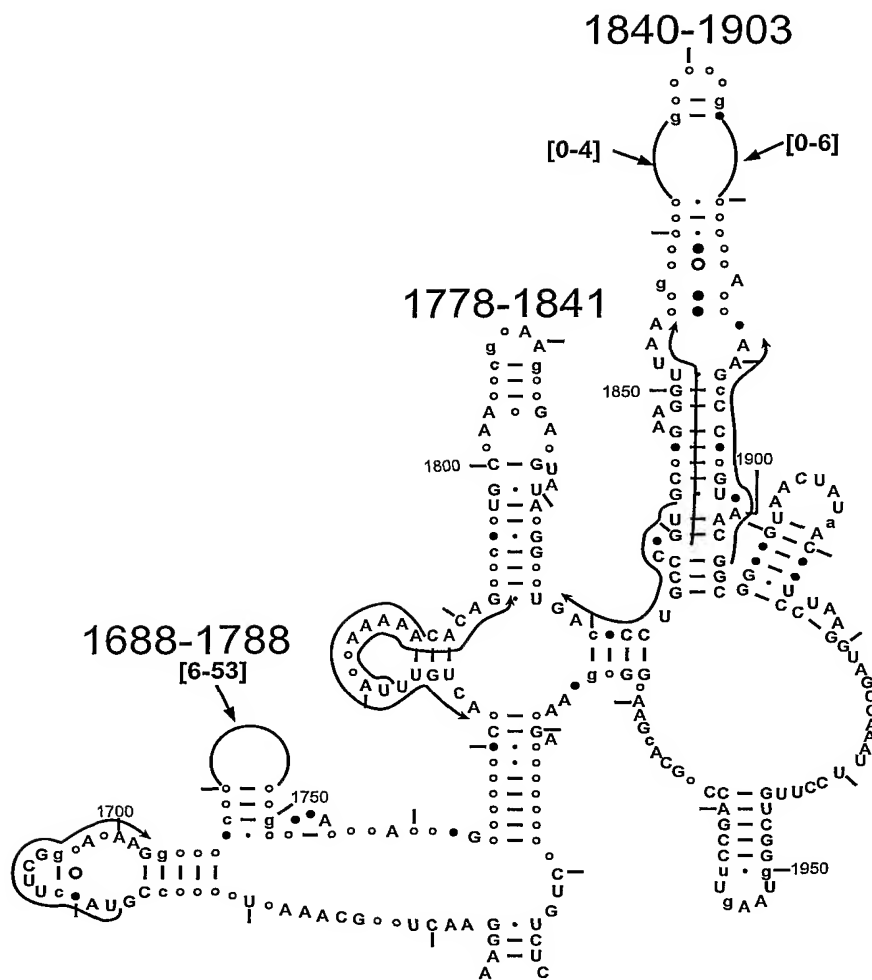
Fig. 1H

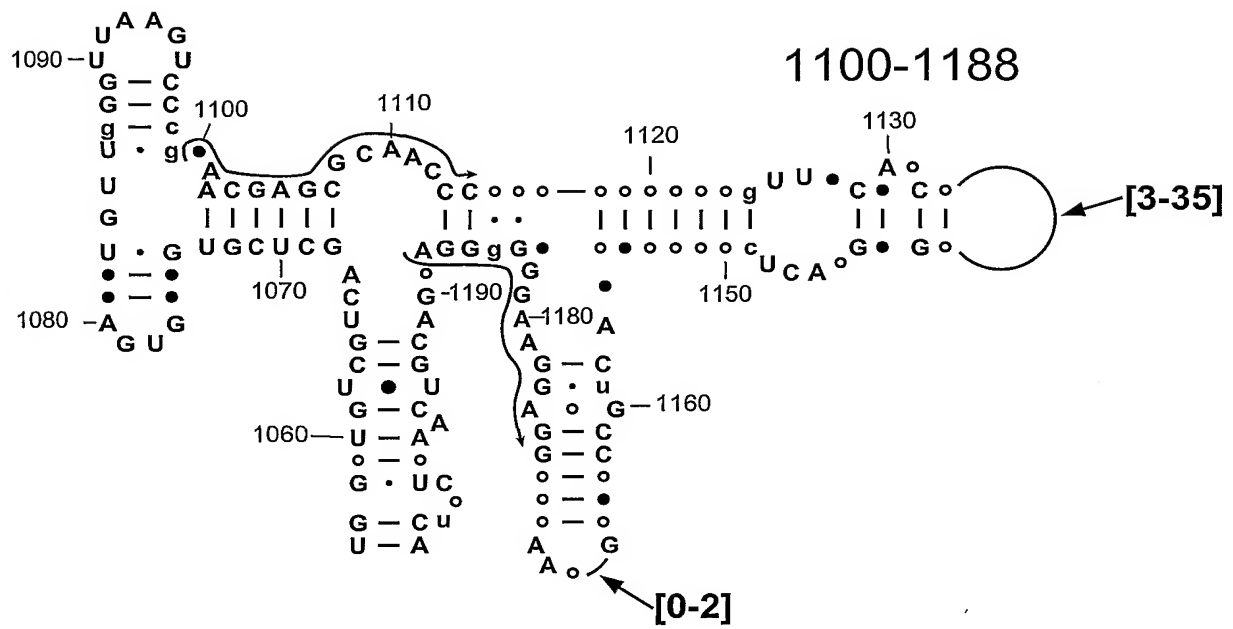
Fig. 2

Fig. 3

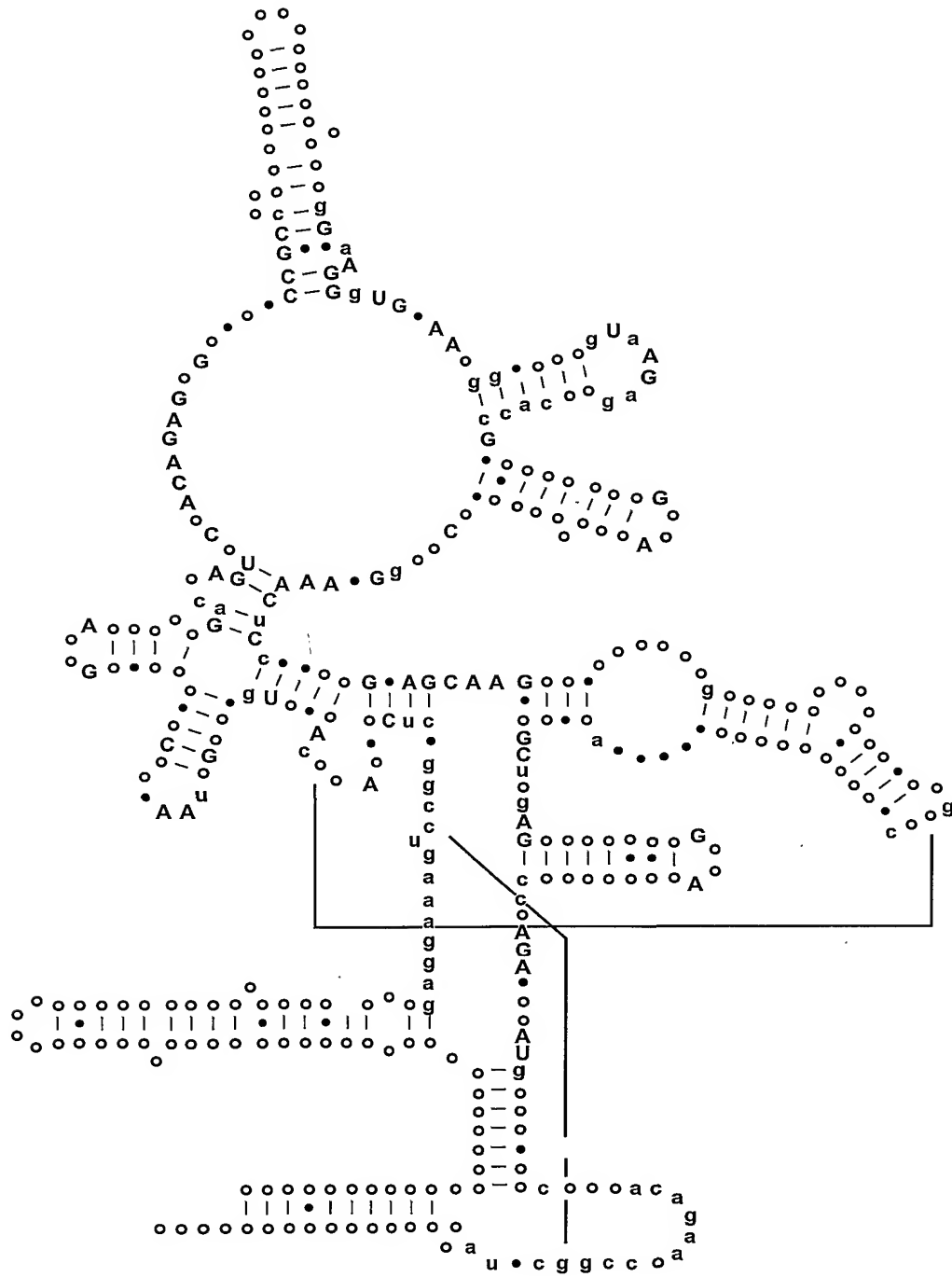


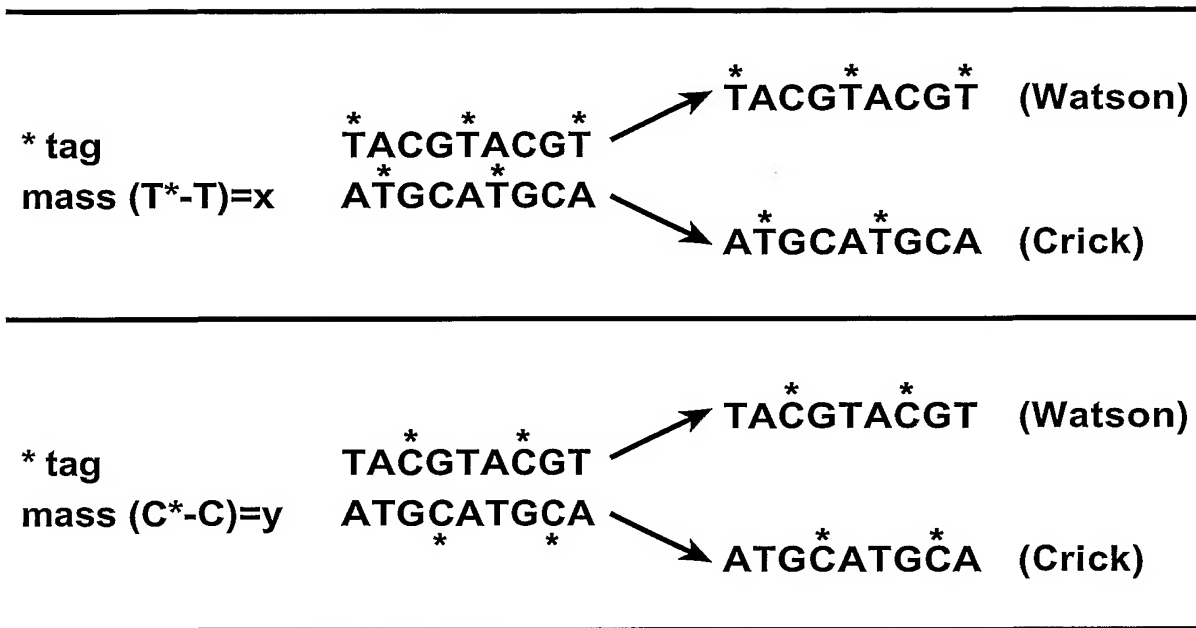
Fig. 4

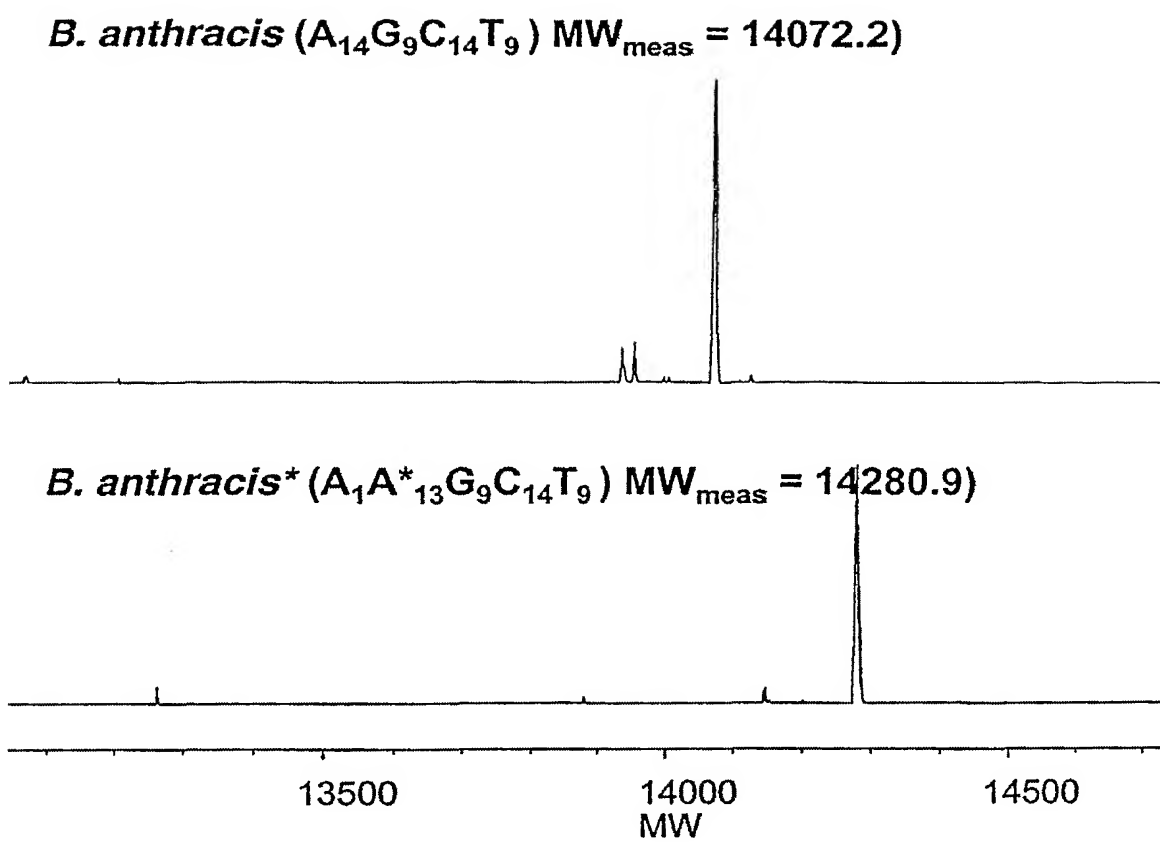
Fig. 5

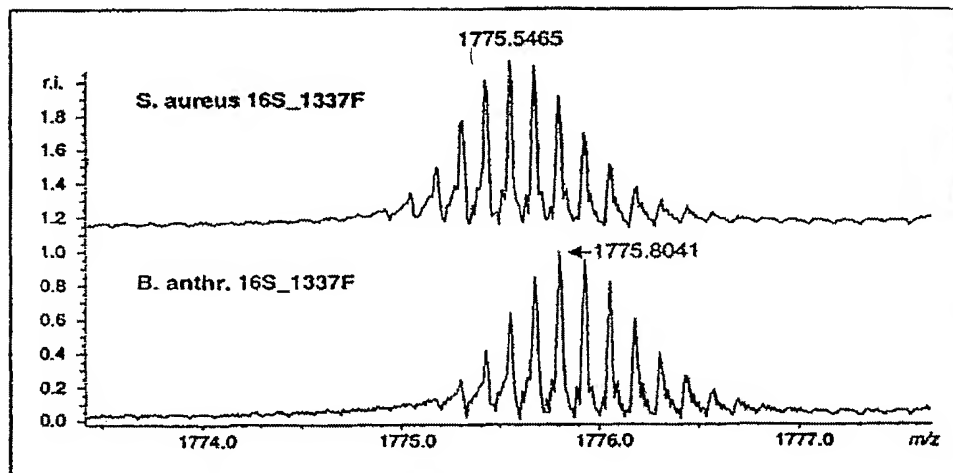
Fig. 6

Fig. 7

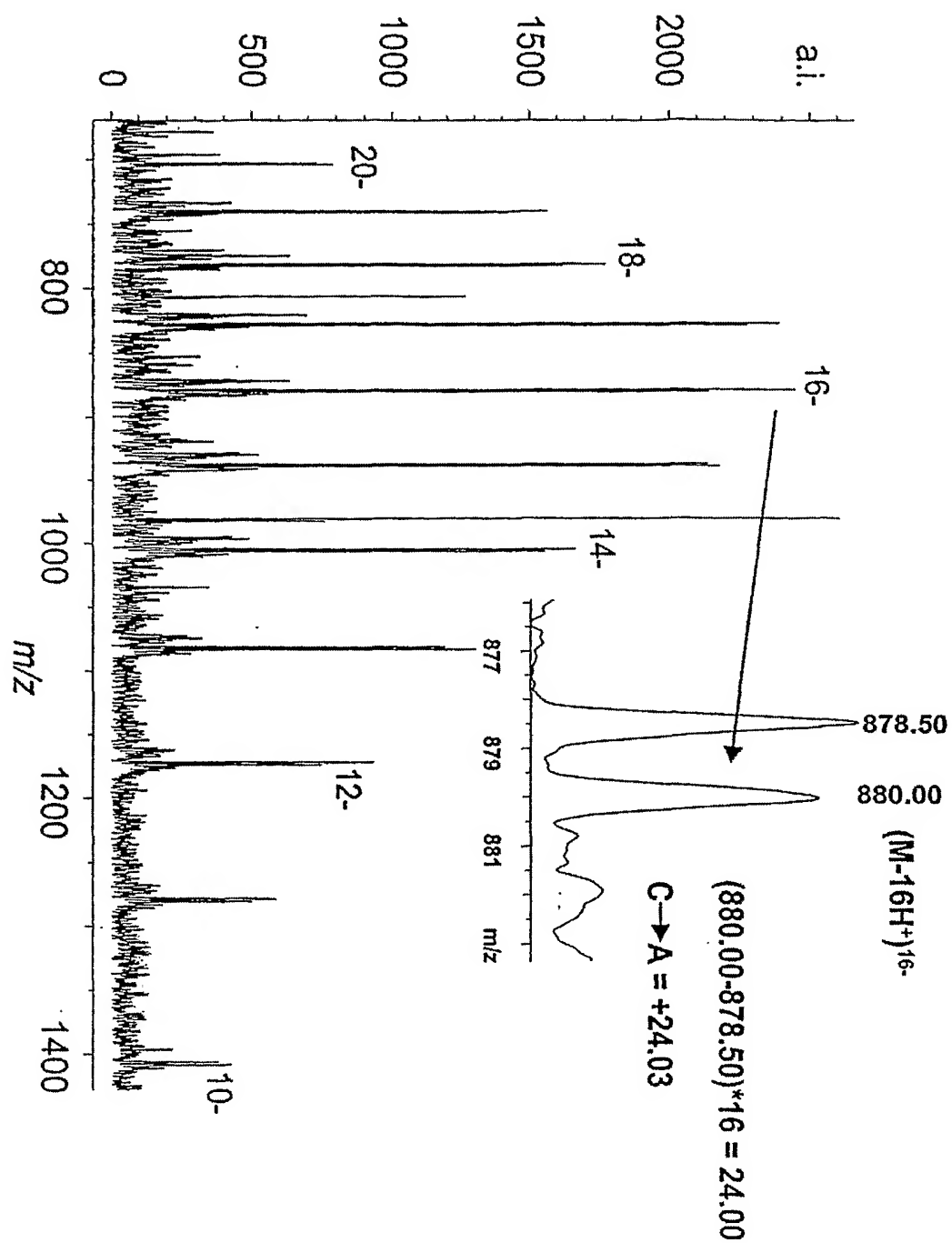


Fig. 8

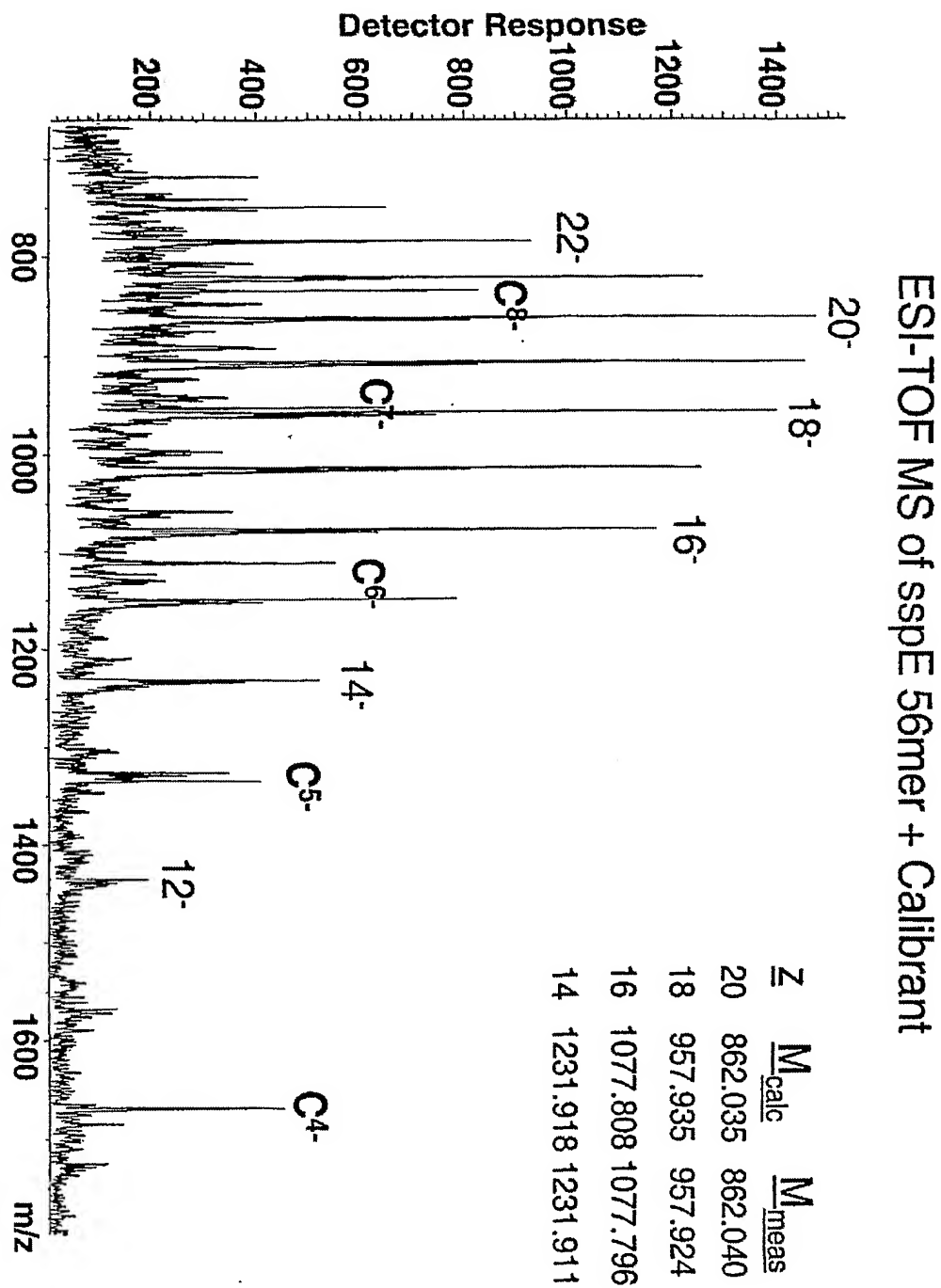


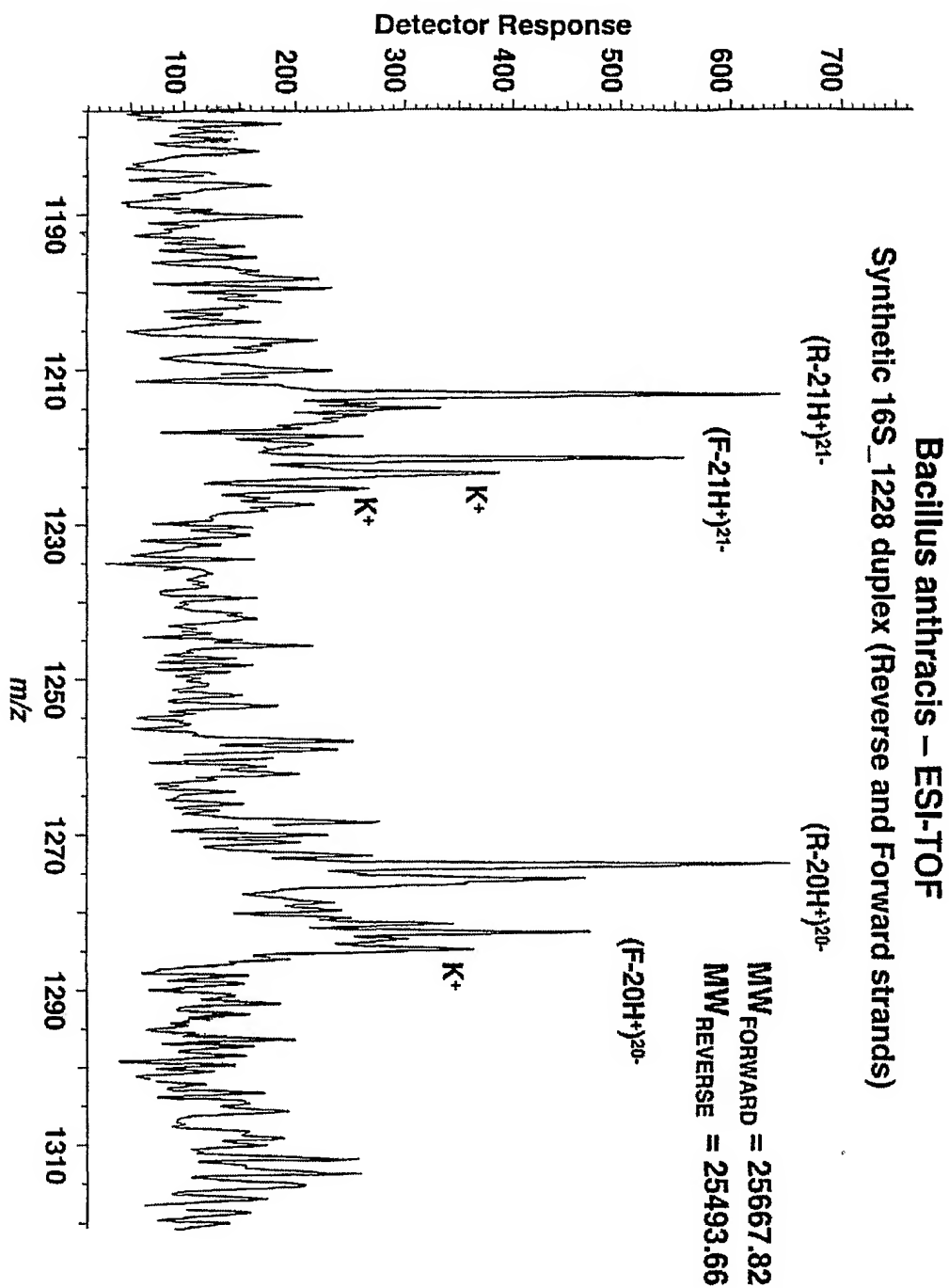
Fig. 9

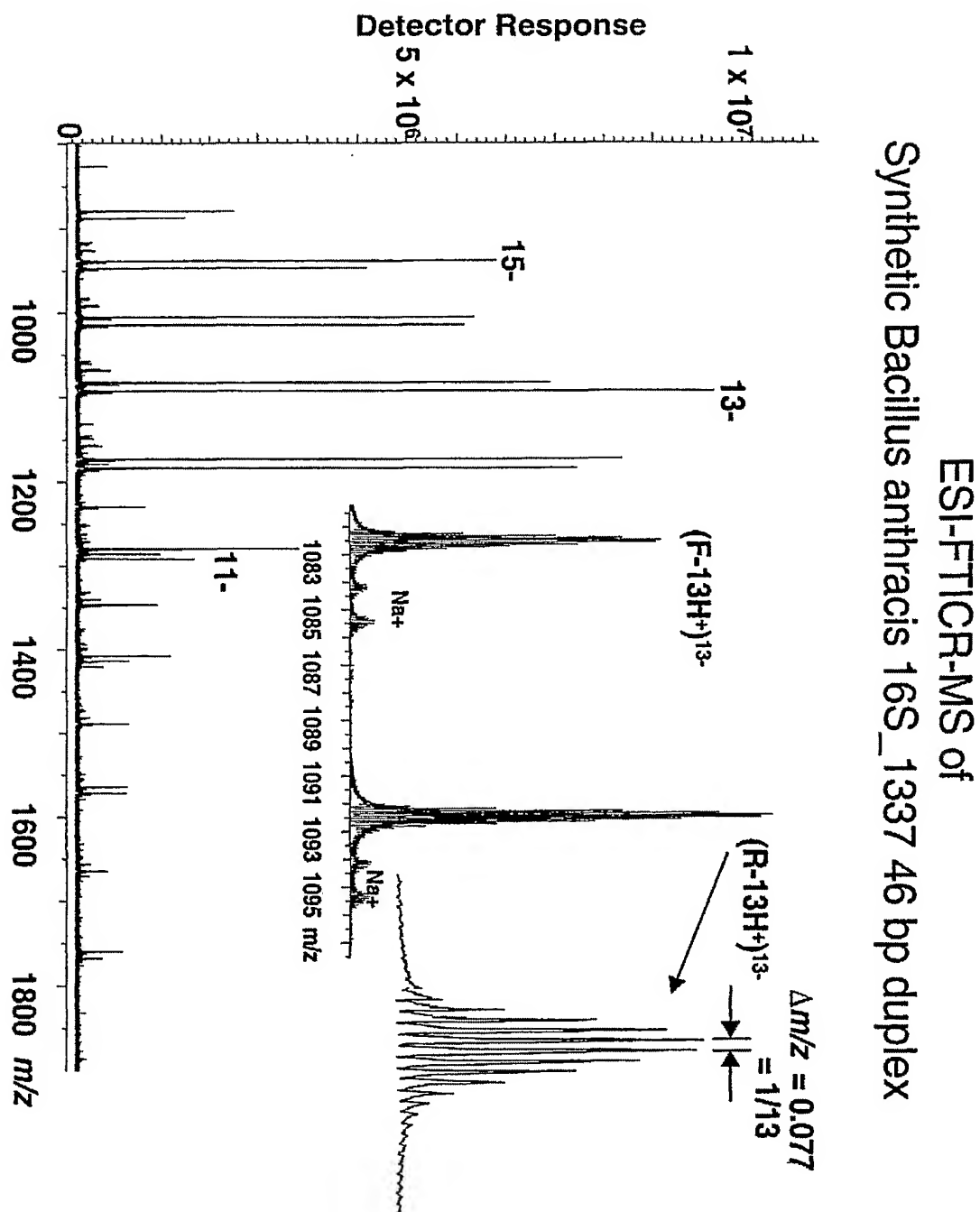
Fig. 10

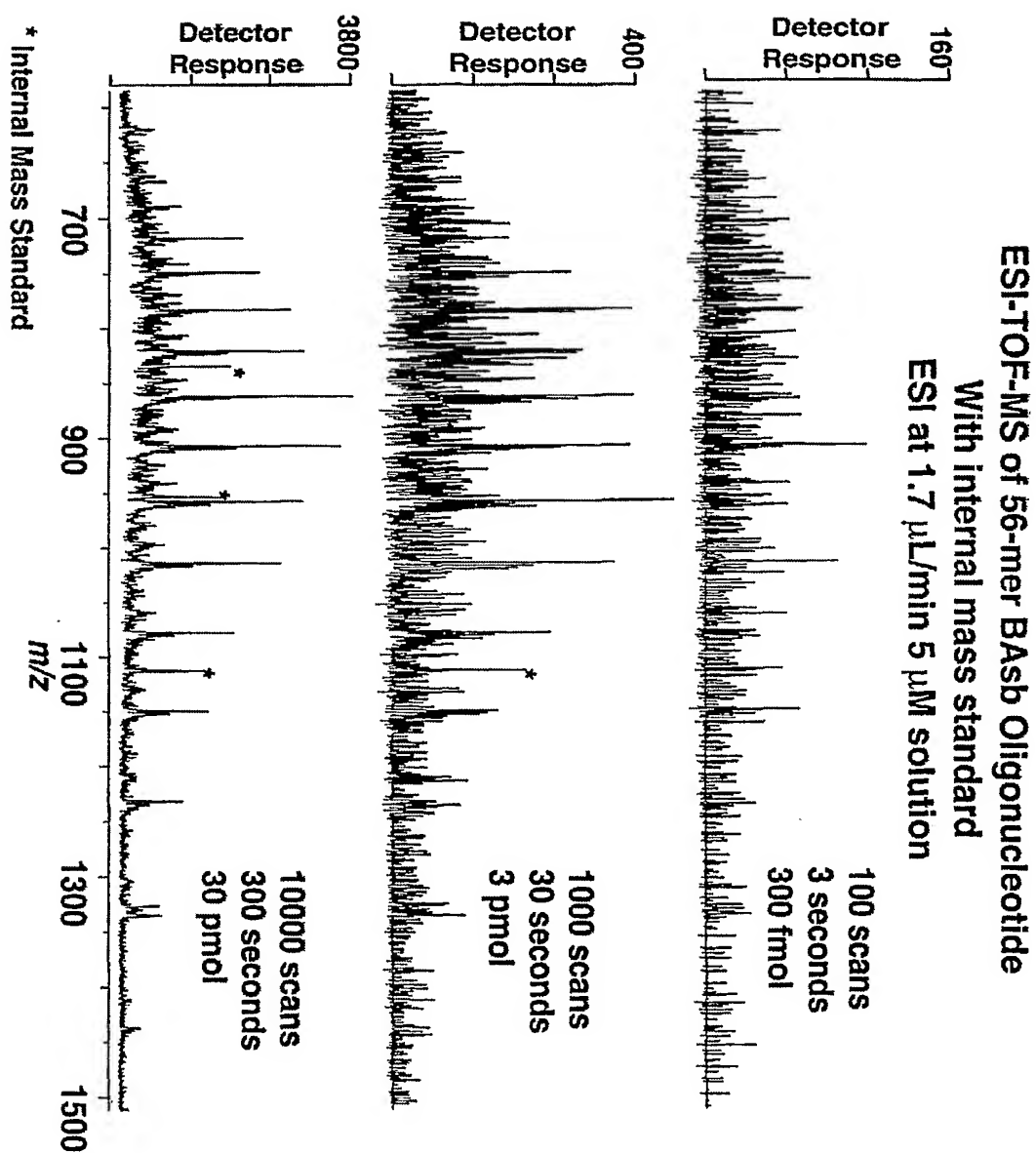
Fig. 11

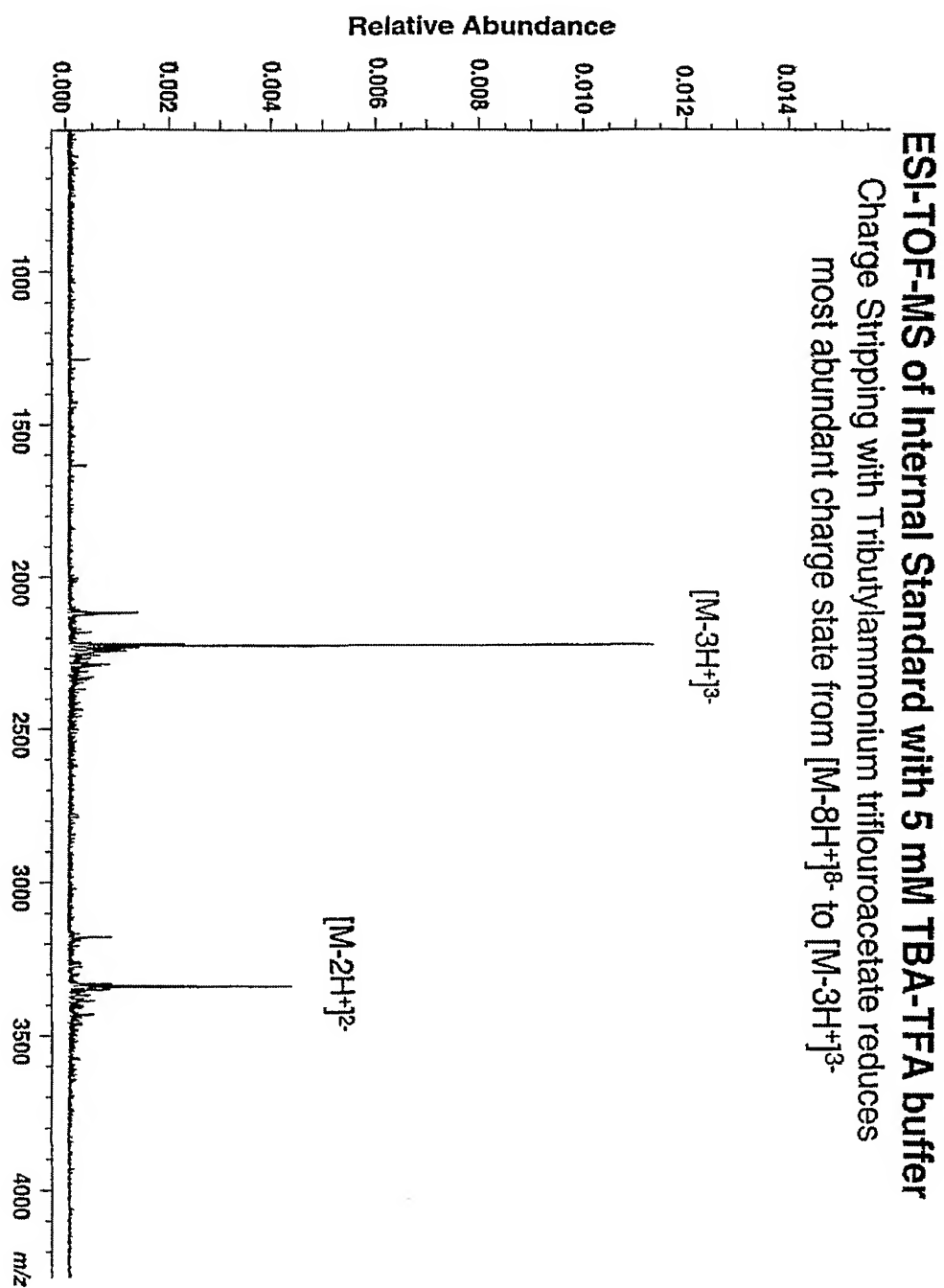
Fig. 12

Fig. 13

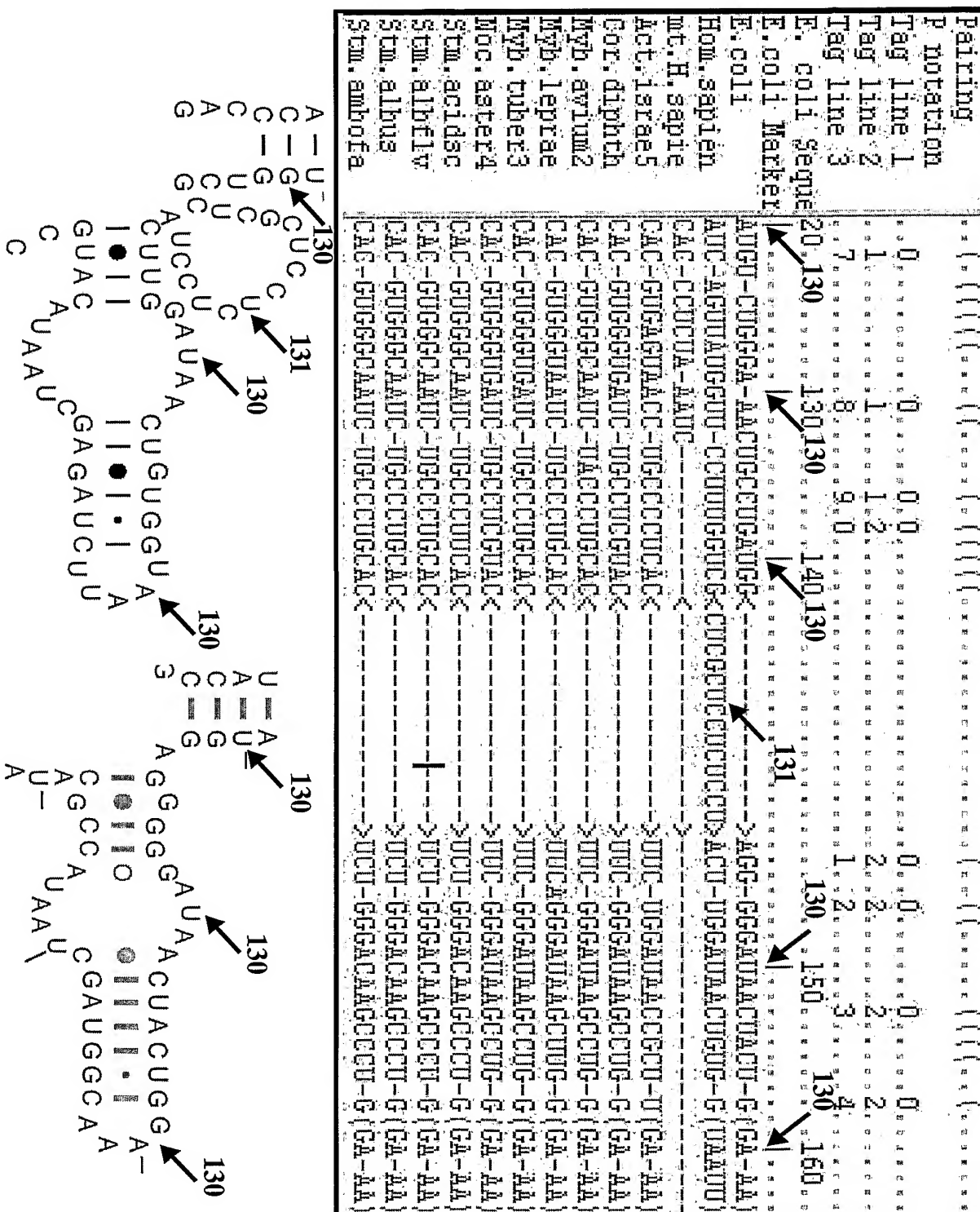


Fig. 14

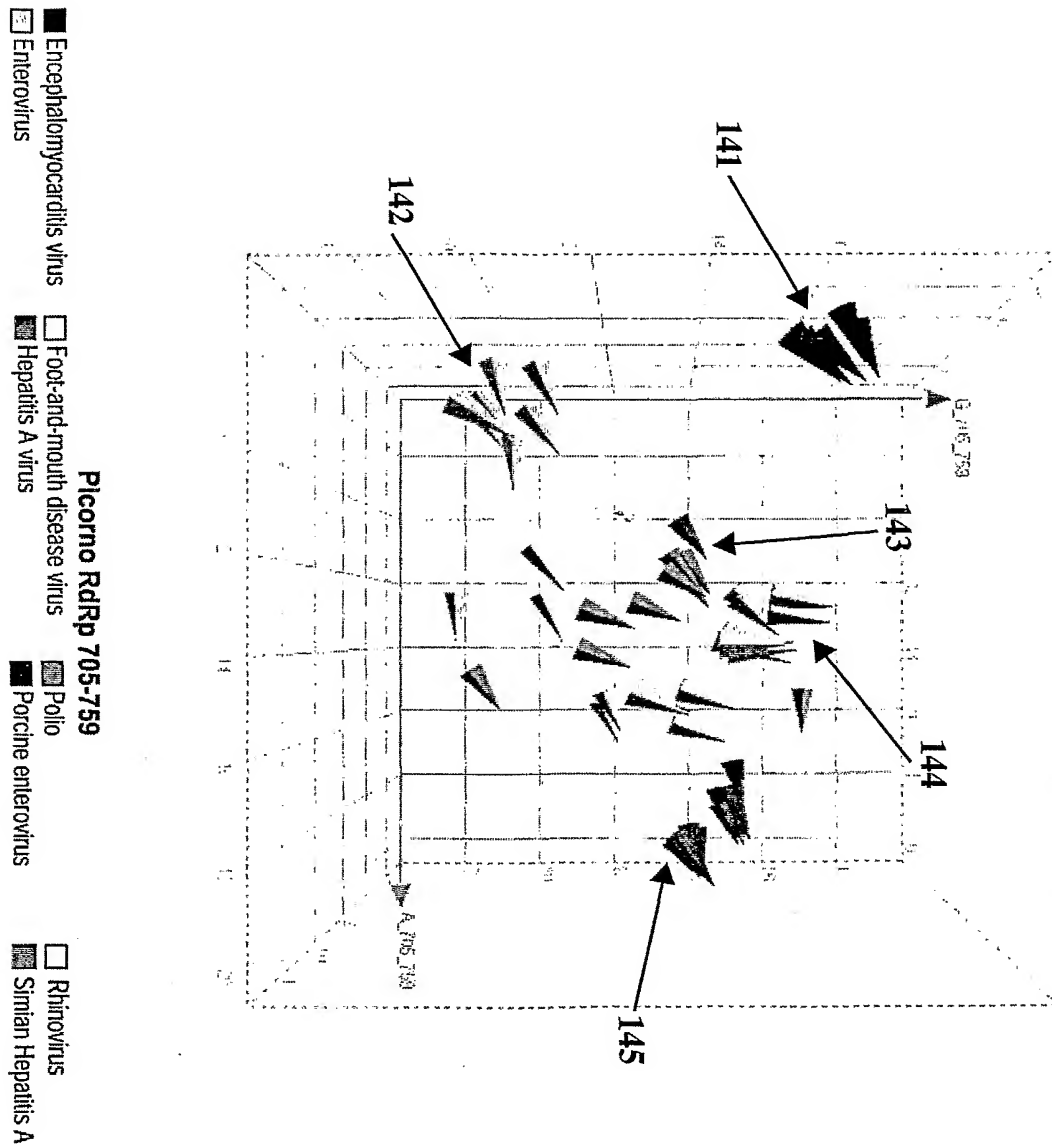


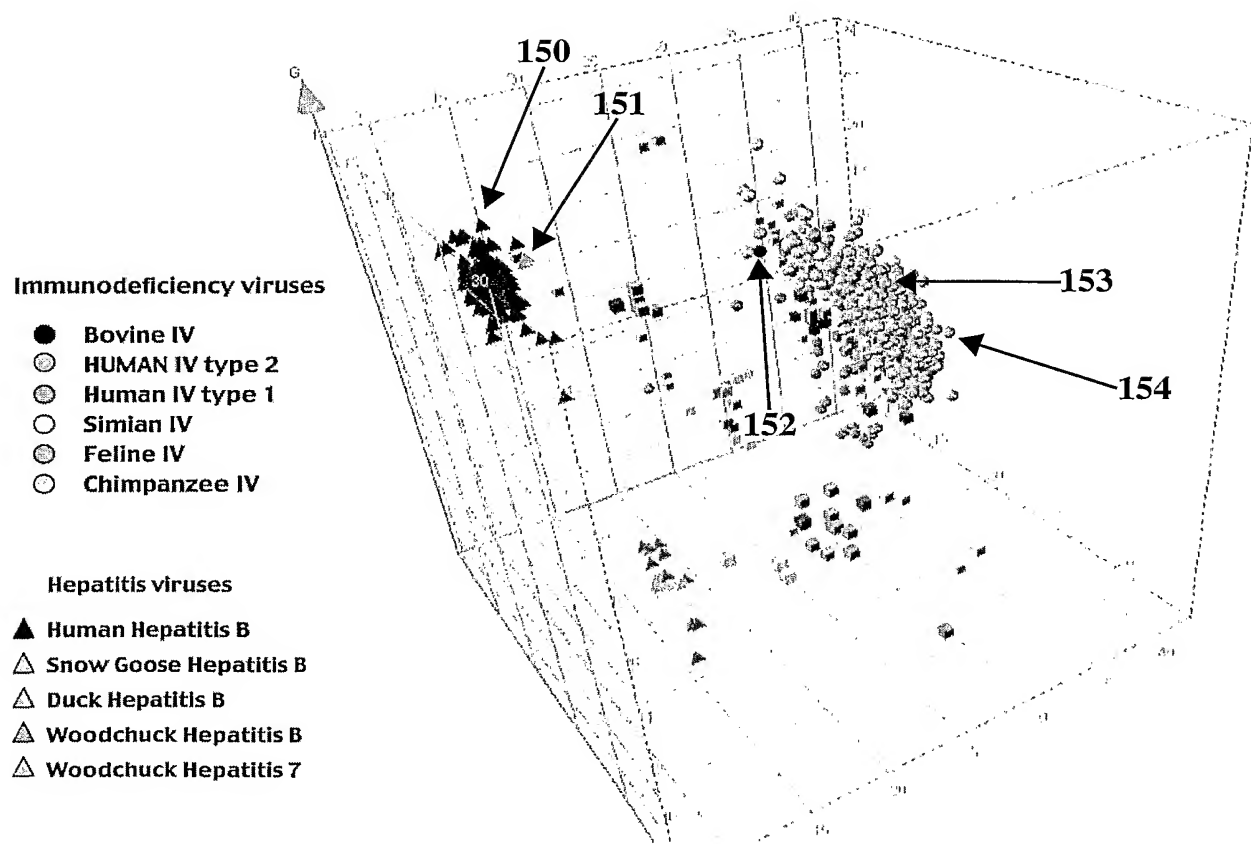
Fig. 15

Fig. 16

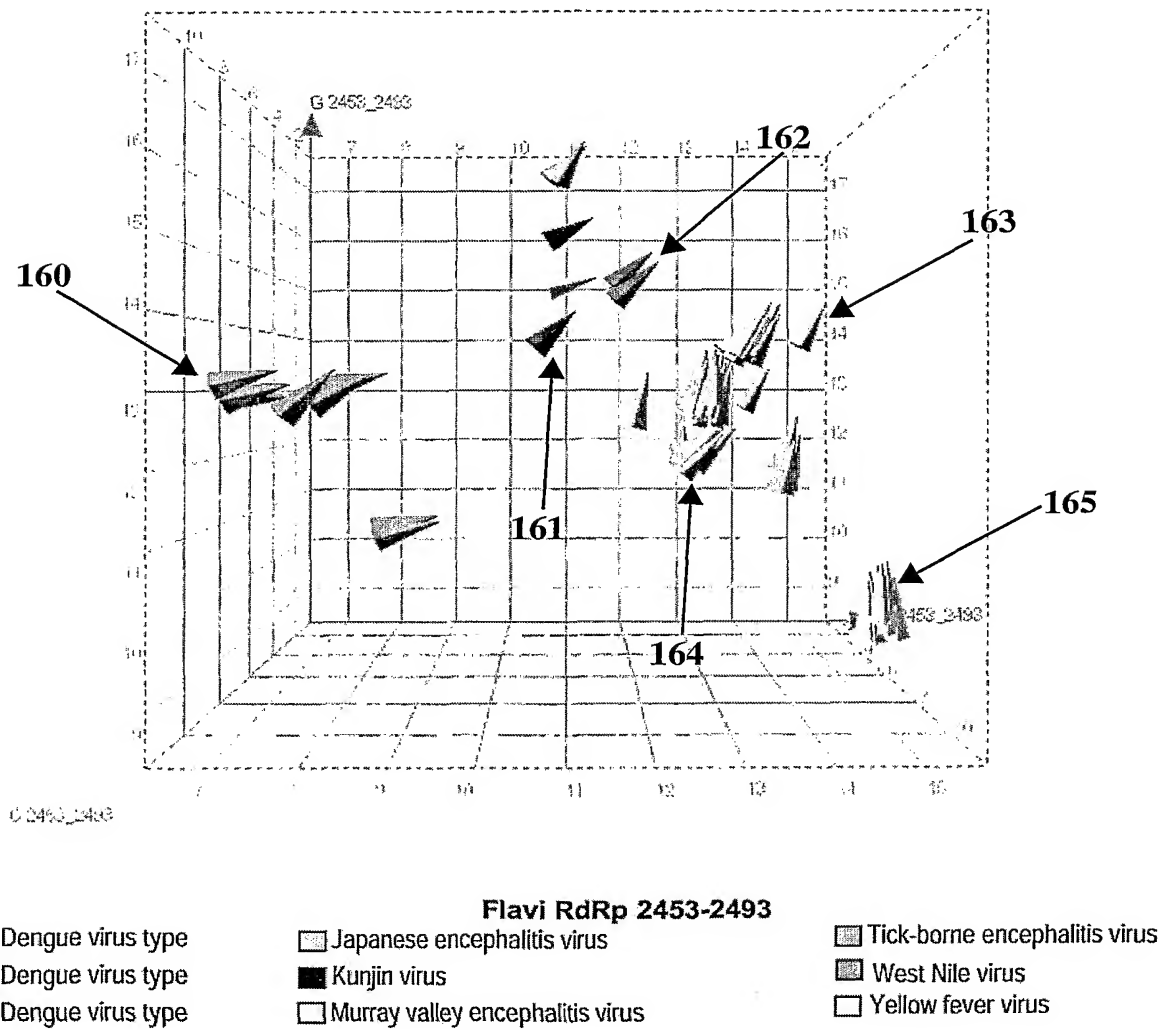


Fig. 17

